

Rebuttal Report

Review of Principal Components Analysis of Data and Review of Inferences about Presence of Biomarkers in the Population of Animals from the Illinois River Watershed

Prepared for:

Tyson Foods, Inc.
Tyson Poultry, Inc.
Tyson Chicken, Inc.
Cobb-Vantress, Inc.
Cal-Maine Foods, Inc.
Cal-Maine Farms, Inc.
Cargill, Inc.
Cargill Turkey Production, LLC
George's, Inc.
George's Farms, Inc.
Peterson Farms, Inc.
Simmons Foods, Inc.
Willow Brook Farms, Inc.

Prepared by:

Charles D. Cowan, Ph.D.
Analytic Focus LLC
4939 De Zavala Road, Suite 105
San Antonio, TX 78249

November 26, 2008



Charles D. Cowan, Ph.D.

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

PERSONAL SUMMARY

1. My name is Charles Cowan. I reside in San Antonio, TX. I was retained by the defendants to provide an opinion regarding the use of principal components analysis by Dr. Olsen for this litigation and the statistical reliability and value of sampling used both by Dr. Olsen and Dr. Harwood. I have personal knowledge of the matters contained in this report.

Education and Experience

2. My background covers 30 years of research and study in the areas of statistics, economics, and their application to business problems. I am Managing Partner of Analytic Focus LLC, a company headquartered in San Antonio, TX and with offices in Birmingham, Alabama and Washington, DC. A portion of our work is conducting research for legal matters, including providing litigation support and expert witness services when requested. Some of our work focuses on measurement and mitigation of risk for financial intermediaries. The final area of our practice is in support of Federal and State agencies needing economic and financial analysis to pursue their missions. Prior to starting Analytic Focus LLC I served as Chief Statistician for the Federal Deposit Insurance Corporation. I was also a Director for Price Waterhouse where I headed the Financial Services Group in the Quantitative Methods Division. I served for 12 years at the U.S. Bureau of the Census where I was responsible for the evaluation of the Decennial Census and held the title of Chief of the Survey Design Branch.

3. I am currently an adjunct professor in the School of Public Health at the University of Alabama – Birmingham (UAB) and previously served as a professor in the Business School at UAB, as a visiting research professor at the University of Illinois, and in other academic and professional positions.

REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

4. A listing of my qualifications as an expert in this case are presented in Appendix 1. My complete resume and a listing of all my publications are presented in Appendix 2. A listing of past cases in which I have been deposed or presented testimony at trial is presented in Appendix 3.

Scope of Assignment & Compensation

5. I was asked to consider the claims made by the plaintiffs in the above referenced case and to offer an opinion on issues pertaining to their claims. This report considers both issues.

Personnel	Fees per Hour
Charles Cowan, Ph.D.	\$425
Senior Financial Analyst	\$395
Senior Research Associate	\$295
Programmer	\$225
Research Analyst	\$125

For expert representation, depositions and testimony, our hourly rate is \$525. Out-of-pocket expenses, including travel, are billed separately and are in addition to the hourly fees.

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

INTRODUCTION

6. I was asked to review the mathematical and statistical foundations for the use of Principal Components Analysis (PCA) in the report by Dr. Olsen and the selection and use of samples by both Dr. Olsen and Dr. Harwood in their reports. In the former case – the PCA – I looked at what a PCA is, how it was used, what methods were employed in actually performing the PCA, and issues in the construction of the data used in the PCA.

7. In the latter case, the sampling, I looked at how the sampling approach was constructed, and the use of samples in drawing inferences. I examine the ability of Dr. Olsen to draw inferences about the sources of constituents of the watershed and the ability of Dr. Harwood to draw inferences about the characteristics of various animal populations from the sample she used.

8. I concentrate first on the report by Dr. Olsen and examine the methodology he employed, and then move to Dr. Harwood's report. The following sections refer to Dr. Olsen's report and address distinct parts of his work:

I. What is a Principal Components Analysis?

II. What Did Dr. Olsen Do?

A. Collection of data from different sources

1. Consolidation of data into a single dataset
2. Conversion of the data into a form suitable for analysis
3. Problems finding data
4. Problems summarizing into averages

B. Missing Data

1. How much?

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

2. Dr. Olsen's Substitutions

3. Problems with data from multiple sources

C. Methods for Treating Missing Data

1. Means

2. Structural relationships

3. Increases in the Variability of the Data

4. Biases in Correlations

D. Non-Detects

1. Use of substitution to allow for non-detects

2. Variability in detection levels

E. Use of Logarithms

1. Comparison of Logarithms to Original Data

2. Potential reasons for use of logs

3. How Logarithms change the relationship studied

4. How the transformation changes the correlation

5. How the transformation affects the non-detects

F. The Number of Principal Components and Rotations

1. Choice of Principal Components for Analysis

2. Use and non-use of rotations for comparative purposes

III. What Did Dr. Harwood Do?

A. General Principals of Sampling

1. What to Measure

2. How Precise?

3. Representativeness

4. How the Samples Were Selected

WHAT IS A PRINCIPAL COMPONENTS ANALYSIS?

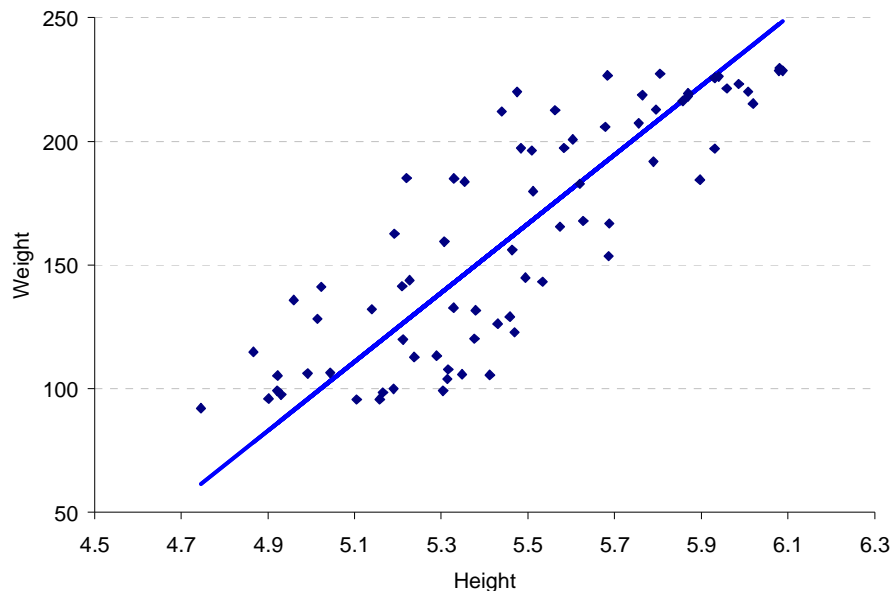
9. PCA is a method used to summarize information. In a research study, a researcher will make multiple observations (collect multiple samples) which contain measures on a number of different factors of interest in the study (variables). A common example is taking measurements of weight, height, girth, body mass index, and similar variables on people.

10. The researcher measures multiple variables quantifying characteristics of the sampled item. These variables will be correlated with one another to varying degrees, and so although multiple variables are collected, there will be less real information available to the researcher because of redundancy between the variables. Principal components is a method for examining and summarizing the amount of information actually collected. A simple example follows.

11. Suppose we have the height and weight of a sample of a number of adults. We know these values are related, and if we chart the values we see that there is a distinct pattern to the data. In Chart 1, as height increases, weight increases. However, we could have also turned the chart around and observed that, as weight increases, height increases. The two values are strongly related, but one cannot say that one causes the other – they just increase together. The straight line that runs through the points is the first principal component – it is the line obtained by minimizing the distance from each point to the line, measured at right angles to the line. Not up-down, not left to right, but the shortest distance to the line for each point. This line measures the relationship summarized in both height and weight, although we do not know what this relationship is. We can call it “size” since that seems to be what it is measuring. It is also the line that captures the most variability in both variables. If the variability for height and weight separately are large, it is now summarized in one variable instead of two so that only the new variable (size) has all the variability.

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Chart 1: Size Measured as Height and Weight



12. A principal component measures a summary relationship between all variables being analyzed, and does so by describing a line that encapsulates the relationship. The line is written as the sum

of each variable, with a weight on each variable to indicate how much it contributes to the relationship.

13. The line above would be summarized as:

$$(\text{Principal Component 1}) = a_1 * \text{Height} + b_1 * \text{Weight}$$

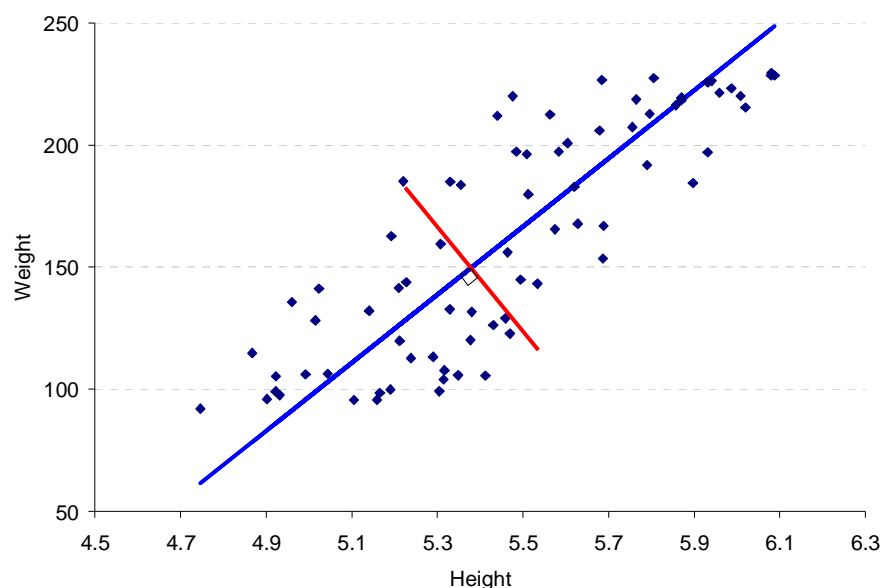
14. We don't know what this value measures – it is an artificial construct based on the relationship between height and weight. We can imagine an underlying relationship in people called “Size” and that Height and Weight are different manifestations of this value. If we want to summarize the set of measurements in one dimension, we have a single variable we measure called size rather than two variables we measure, like weight and height separately. There is now one dimension instead of two and we have eliminated redundancy in the data. This doesn't seem important in the case of two variables, but with multiple variables, each measuring only a piece of the underlying factor, this can be very important.

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

15. The origins of principal components can be found in psychology and education, and are the foundation for IQ tests and educational attainment tests. In education, we don't have a single variable that measures everything we know – it's tested by asking multiple questions and then obtaining a final score on a particular subject. The Scholastic Aptitude Tests (SAT) used as part of the application to get into college are done in this way, testing how much a person knows in a subject. Not all the questions get the same weight – easier questions get less weight than hard ones, and the weights to combine all questions are computed using a technique like Principal Components Analysis.

16. There are as many principal components as there are original variables. The second, third, fourth, and so on principal components are measured at right angles to earlier principal components. Each new principal component captures what is left over of the variability in the data from the previous principal components.

Chart 2: A Second Principal Component



17. On Chart 1 a second principal component could be measured at right angles to the first one. This is done on Chart 2. The second component measures a different relationship between height \ weight.

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

18. The second principal component is a line at right angles to the first. written as

$$\text{Principal Component 2} = a_2 * \text{Height} + b_2 * \text{Weight}$$

19. This second relationship is uncorrelated with the first principal component. It summarizes variance left over. In Chart 2, the line is much shorter than the first principal component because there is less variability left over to explain. Note that the later principal components may or may not measure something of value – they could just be measuring whatever leftover variability there is in the data.

20. Or they could be measuring a unique basis for why the first few principal components do not fully explain the data. Continuing the height and weight example, the further a point is above the first principal component (in blue), the more overweight the person is **relative to the norm defined by the first principal component**. Points below the first principal component are people who are underweight **relative to the norm**. In this example, the main principal component establishes a norm for the relationship of height and weight. The second principal component measures how far one is above or below the norm – much as a physician would decide whether a patient is overweight or underweight.

21. There are four other issues to understand about Principal Components Analysis. These relate to strength of relationship, sampling, interpretation, and utility.

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Strength of Relationship

22. If the distribution of the (height, weight) points is very close to the line fit through them (the example above was Principal Component 1 = $a_1 \cdot \text{Height} + b_1 \cdot \text{Weight}$) then the relationship is very strong. If the distribution of the points is wide and not close to the line, then the relationship is weak. Each variable contributes “one” to the overall variability. With 26 variables, this means that the overall variability that can be explained is 26.

23. When a solution is found in Principal Components Analysis, each principal component (called an “eigenvector”) has a corresponding measure of how much variability is explained (called an “eigenvalue”)¹. The eigenvector is the set of weights applied to the variables. In the relationship [Principal Component 1 = $a_1 \cdot \text{Height} + b_1 \cdot \text{Weight}$], the values (a_1, b_1) together are the first eigenvector (a vector is a collection of weights in an equation for a straight line).

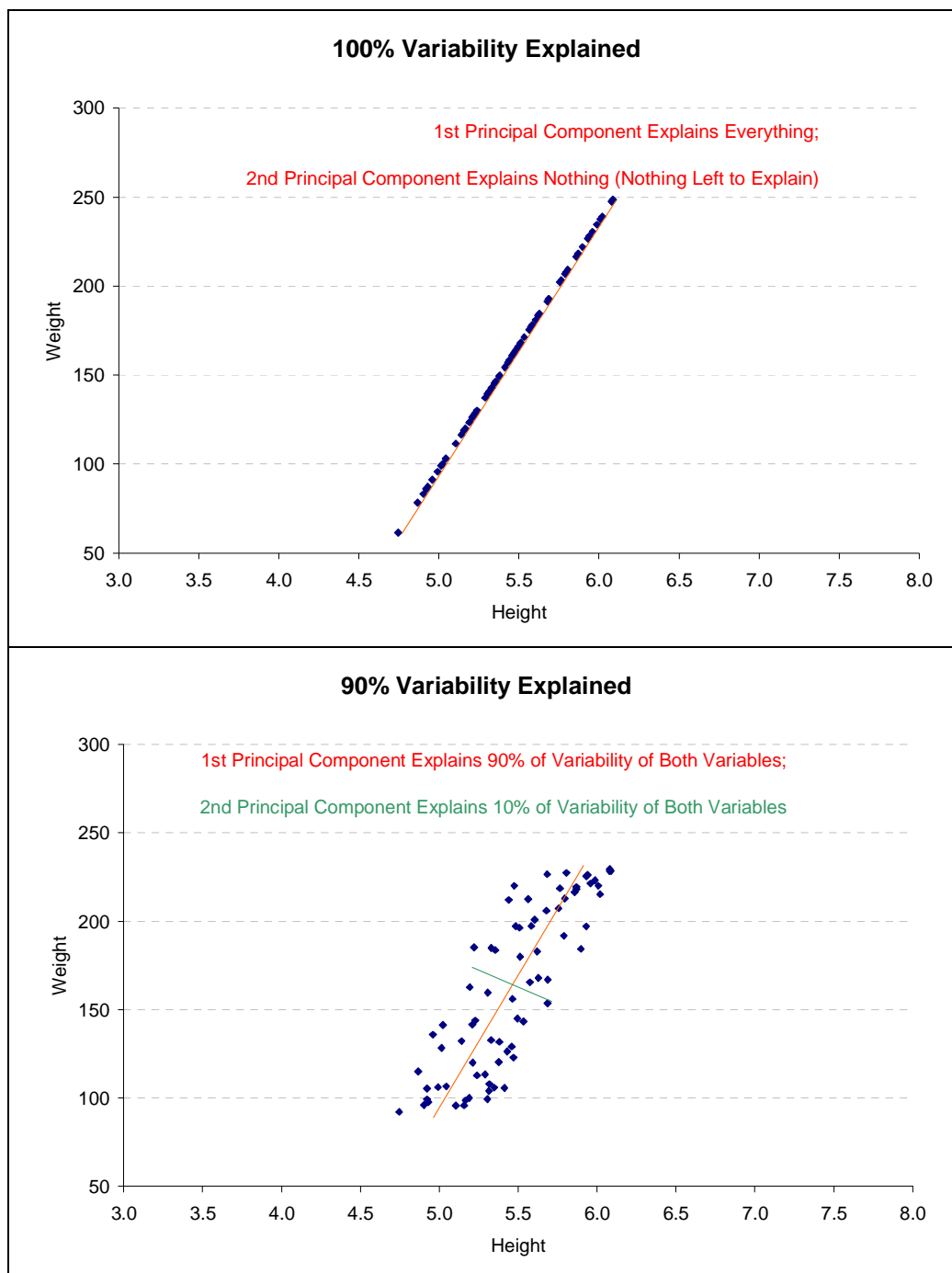
24. The eigenvalue (λ_1) for the first vector (a_1, b_1) is a measure of how much overall variability in **all** the variables is explained. The values of (a_1, b_1) are chosen so as to make λ_1 as large as possible. In other words, the line is the best fit – regardless of how strong or weak the relationship is – to all the points because it explains the most variability. That doesn’t mean it does a great job of explaining the variability. Rather, it’s the best we can do given how strong the relationships are. Chart 3 gives four different relationships from the same example where there is a perfect, a strong, a moderate, and no relationship.

¹ "Eigen" is German, meaning “inborn or forming a natural or inseparable part or quality of”, from Dictionary.com.

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

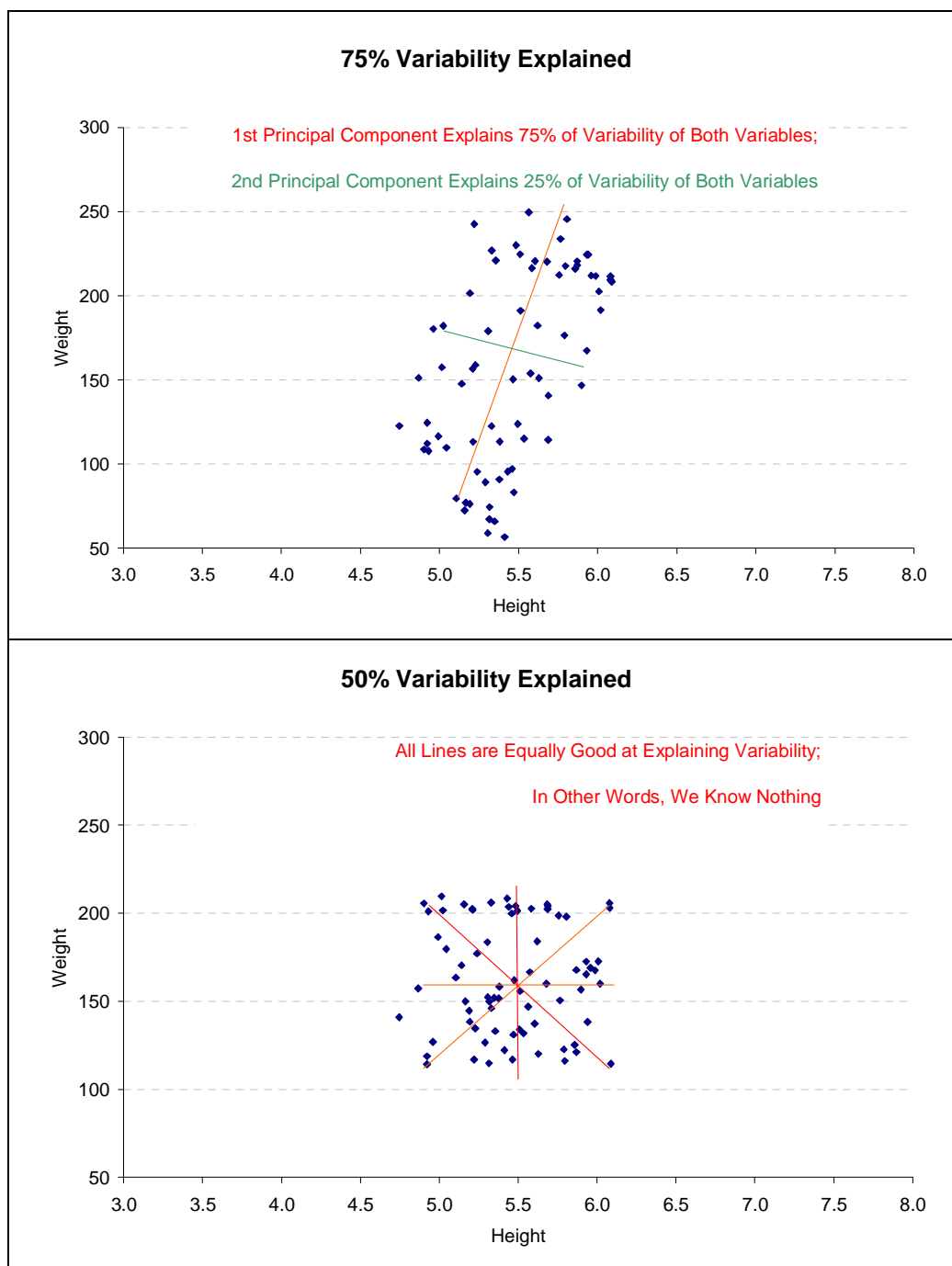
Chart 3: Strength of Relationship Examples

25. The eigenvalues for the next four charts are 2 (out of 2 = 100%), 1.8 (out of 2 = 90%), 1.5 (out of 2 = 75%), and 1 (out of 2 = 50%).



REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED



26. In the bottom chart, the 50% Variability Explained means that Principal Component 1 and Principal Component 2 each explains the same amount – in other words, there is no advantage or new information in the principal components since they don't explain or account for any more variability than the original two variables.

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Sampling

27. It should be obvious, but it needs to be said: any statistical technique is only as good as the data collected. In particular, if the PCA is based on a sample, then to be able to say something about a population, one needs to have a projectable sample. A projectable sample is one where the methods used to select the sample enable the researcher to extrapolate from the sample to the population. For example, a sample of voters, if selected correctly, can be used to project to the population to forecast the outcome of an election. A group of voters who respond to a CNN on-line poll is NOT a random subset of the population and is meaningless for use in determining what voters in the population are thinking.

28. If Dr. Olsen's sample is not projectable to a broader population or to the area covered in his analysis, then the PCA has no worth in making a statement about what is occurring in the Illinois River Basin. Other reports delve into the quality of the sampling. If it is established that Dr. Olsen's sampling is not representative of the Basin or is biased in some fashion, then the PCA he conducted has no determinative value.

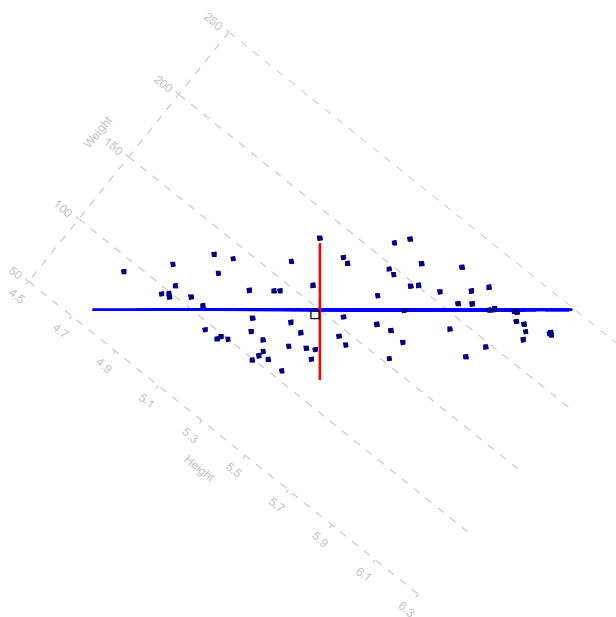
Interpretation

29. When we measure a dimension in a PCA, the question arises as to how to measure the new dimension. In other words, what is a large or small value? In the original data, larger values of height and weight are easy to determine, but that is because they are measured separately on the horizontal and vertical dimensions. The principal component in Chart 1 is hard to interpret as it stands because it requires two dimensions to display it.

30. But the principal component is supposed to be only a single (underlying) dimension. So we can turn the chart so that we can use the line as the dimension we want to summarize.

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Chart 4: Rotated Principal Component



31. In chart 4 we can now measure “size” on one dimension, and “off-norm” as it’s own dimension going in a different direction. This is the exact same data and the same lines, but rotated to give meaning to distances right and left and up and down

32. The choice of rotated outcomes is important for the interpretation of the data. Dr. Olsen does not consistently

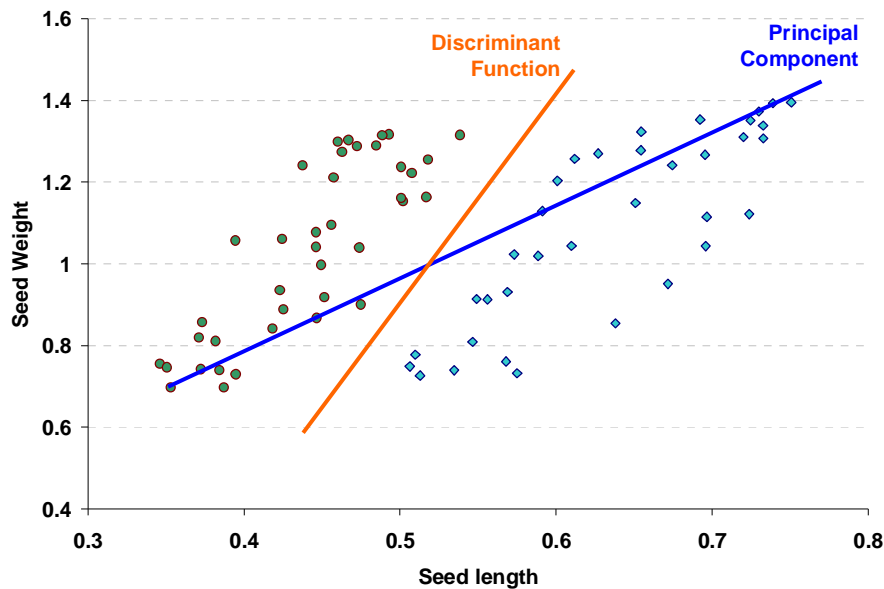
report the rotated values – in fact, he goes from unrotated to rotated solutions without recognizing that there are problems of interpretation. This topic will be discussed later.

Utility

33. Principal components is an excellent technique for discovering a dimension or factor that is continuous and giving values to indicate large or small. It is not usually an appropriate technique for discerning the difference between two groups. Often, a completely different technique should be used to differentiate between two or more groups. In the following example, we have a chart showing weight and seed length for two types of flowers. The principal component can show size, but it can’t be used necessarily to differentiate between the two groups. There are other methods that are much better suited to differentiation. One such technique is discriminant function analysis.

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Chart 5: Principal Components versus Discriminant Functions



34. Suppose a grower was trying to differentiate between two types of seeds for sale, as seen in Chart 5. Using a principal component the grower would be wrong half the time, whereas if the grower used a

discriminant function, he could easily distinguish between the two groups.

35. Discriminant Function Analysis (DFA) allows for tests to determine if it is possible to differentiate between two groups; PCA does not. DFA attempts to maximize the difference between two or more groups; PCA does the reverse. PCA homogenizes the data to get an underlying factor. Dr. Olsen chose a technique that requires a subjective judgment regarding how to divide his data into two groups². There are other techniques like the one demonstrated above that give an objective method for determining if it is possible to distinguish between two groups and the test for determining how to best distinguish between the two.

² The choice for the threshold is 1.3 for his first principal component. CDM Report, page 6-60

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

WHAT DR. OLSEN DID

36. This section reviews the steps taken by Dr. Olsen to create a data base and conduct his analyses.

Collection of data from different sources

37. For the samples taken by a variety of entities, Dr. Olsen had a data base created made available to us in Microsoft Access³. This data combines data collected by the plaintiffs and other samples selected by the USGS⁴. Data is stored as individual observations on specific chemicals or organic matter, identifying the sampling group, the sample within the sampling group, the particular constituent and the amount found in the sample. Different samples had reports on different chemicals or bacteria. There are over 100 indicators measured in the samples, but in each sample there are reports on only some of the 100, so not all samples have measurements on all indicators used in the analysis process.

38. For some samples, there are more than one measurements for a particular indicator (chemical or organic constituents, e.g. bacteria), and so these were averaged in the sample that Dr. Olsen analyzed. Thus, there may be only one observation on aluminum in a sample group, so it is the average. On the other hand, there may be four measurements on fecal coliforms, and the value used from the sample is the average from the four observations⁵.

³ CDM Report, Section 4, Database Compilation and Maintenance, page 4-1

⁴ CDM Report, Section 2.10, USGS Sampling, page 2-39 and USGS in DB page 6-38

⁵ "EDAnalyzer also has an option for creating (or averaging) the cross-tabulation by sample or by location; e.g., in the case of by location, the data for a particular variable with multiple samples assigned to that location would be averaged during creation of the cross-tabulation", CDM Report, page 6-47

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

39. **This is the first key problem in Dr. Olsen's analysis.** He has samples of different sizes summarized as a single observation for his analysis, and thus all data in the analysis are treated as if it has equal contributions to the variability in the data. The truth is, the variability of the data for the bacteria is much greater than the variability for the remaining 22 variables. The reason that the bacteria are more variable is that there are many more observations for bacteria, and these have additional variability when averaged. He summarizes the bacteria data in one point rather than keeping the distribution of bacteria values. **This makes it seem that the bacteria is less variable than it actually is, because the average must be less variable than the original set of multiple observations (a basic principle of statistics).** **If the real variability of the measured data were represented in the summary database used for the PCA, the relationships between the bacteria and all other values would be greatly different, and the results of the analysis would be greatly different.**

40. As it stands, **Dr. Olsen does not retain or analyze a principal component that summarizes the bacteria – he throws it away.** If the bacteria had the correct variability represented, inclusion of this variability would cause the results to be greatly different. Dr. Olsen did not analyze the variability in the data, though he offers that he has. He has disguised the variability through the averaging process, thus **giving too much weight to some variables like phosphorus and not enough weight to other variables like the bacteria.**

41. A second outcome results from this flaw – one that is even worse. Since all of Dr. Olsen's principal components are derived from summary measures of variability, **the correct calculation of variability would change all of the weights he derived and completely change the outcomes that he presents, and change them in ways we cannot project.** This invalidates any of the results Dr. Olsen submits from the Principal Components Analysis.

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Replication of Dr. Olsen's Analysis Dataset

42. Basically, we can't. Although we followed the paradigm described above on all the Access data sources presented to us (specifically Access database 20), we could not replicate the initial analysis dataset (the Excel subdatabase SW3) exactly. The sequence of construction is to convert the Access database to a large Excel database to a summary extract used for the actual PCA. However, **it is not possible** to go from the Access database, which is the original repository database, to the database used for the PCA analysis. For most observations, we can replicate the outcomes exactly. With some additional guesses as to the treatment of unusual observations, we were able to replicate more. But there are still a number of observations where we cannot exactly replicate the data that Dr. Olsen analyzed. A more complete description of how observations were replicated is given later in the report.

43. **This is the second key problem in Dr. Olsen's analysis.** Dr. Olsen or his colleagues were **inconsistent in their treatment of the original observations** to obtain the dataset they analyzed. There is random noise or possibly bias introduced in the data he analyzed since he followed different procedures for different observations. Because of this, the data he analyzes represents different things since the different treatments mean that not all measurements are measuring the same thing. Further, a real scientific study should be able to be replicated by another scientist following the procedures of the first. We cannot – there is no way to take a single set of procedures, either as outlined by Dr. Olsen or modified through detective work, to obtain the dataset that Dr. Olsen analyzed.

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Missing Data

44. As noted above, not all samples have measurements on all observations. In fact, there is a very significant amount of missing data. Dr. Olsen disguises this by substituting for the missing data. He plugs in the mean of a variable for the actual (though missing) value. Only 267 of the 573 samples used by Dr. Olsen have complete data. This means only 47% – less than half – of the observations have real data actually observed in the field. This means that more than half of Dr. Olsen's observations have data that Dr. Olsen substituted rather than real data.

45. **This is the third key problem in Dr. Olsen's analysis.** Dr. Olsen has plugged in so many missing values that a very significant part of the dataset is **made up** by Dr. Olsen. While he analyzes both the data set with no records with missing data and a second data set with substituted data, he fails to admit that he has plugged in values that skew the correlational structure. Dr. Olsen substitutes the mean for a missing value⁶: if aluminum is missing, he substitutes the mean for aluminum from the other sampling sites where aluminum was recorded. This means that he can take data from sites that are in his view poultry impacted and substitute this data into a site that he would not consider being poultry impacted, completely skewing the dataset to show what he wants to show.

46. Dr. Olsen was also missing data in a second way. Samples selected by the USGS measured some different values in the chemicals or processed and tested them in different ways. In particular, total dissolved solids, Sulfate, Total Kjeldahl Nitrogen (TKN) and all three measures of phosphorus were all analyzed in a different way⁷. There are other differences

⁶ There is no direct statement in the CDM Report that states missing values are replaced with means but replacing missing values with means is the only way to reproduce the results from Dr. Olsen's analysis.

⁷ CDM Report, page 6-36

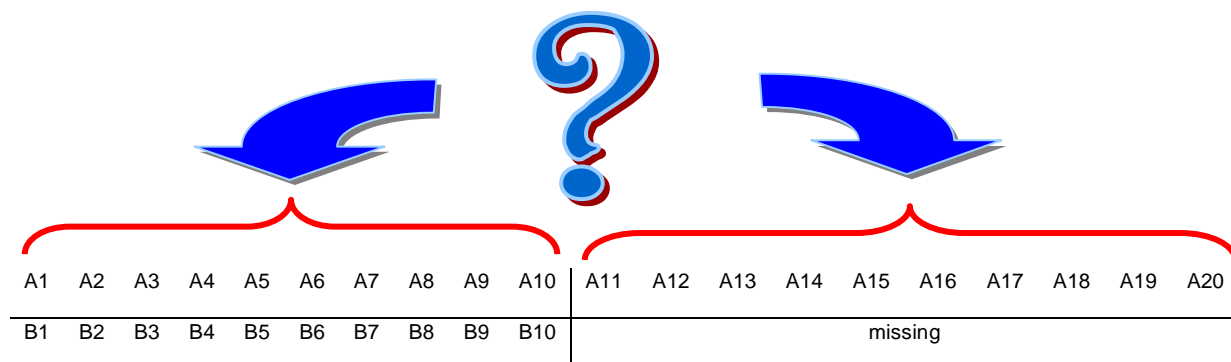
REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

between the USGS data and the remaining data, many of which may be even more substantial in their impact (for example, flow rates differed in the two sets of data). Dr. Olsen doesn't conduct any test for the implication this might have on the PCA. Such a test is presented later in this report.

47. This is the fourth key problem in Dr. Olsen's analysis. Dr. Olsen used data from two sources as if they were equivalent, without testing to see if they measured the same outcomes.

This is completely contrary to scientific method, and in particular is a procedure that undercuts the analysis Dr. Olsen is trying to perform since it is another source of variability in the data.

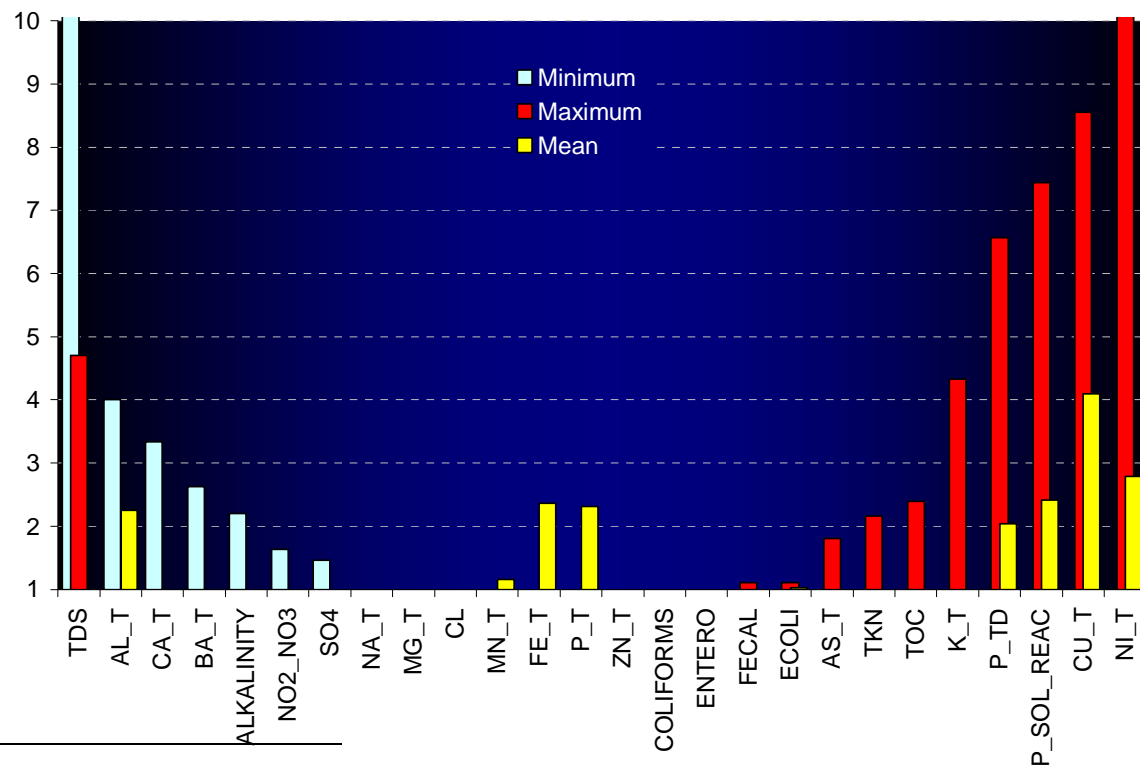
48. Finally, in looking at observations with and without missing data, if the data were just missing at random, we would expect that values in cases with missing data would be just like values in cases without missing data. Suppose we have only two variables: A, and B. Variable A has all of its observations, variable B is missing half of its values. We divide the data into two sets: observations that have measurements on both A and B, and observations that have measurements on only A. If the measurements on B are missing at random, then we would expect values in the first half of A to be like observations in the second half of A.



REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

49. When we examine the data from Dr. Olsen's files, we find this isn't even remotely true for the 26 measures he uses⁸. We looked at the means from the 267 observations that had no missing data. We compared these to the means from the 306 observations that had some missing data. We compared these to the means from the 306 observations that had some missing data. Because the 26 different variables have different scales, we took the ratio of the mean from the first group (no missing) to the mean of the second group (some missing). If the means were the same, this ratio would be unity (1.0). If the ratio was between zero and one, we inverted the value so that it would be measured on the same scale from one to infinity. We performed the same operations for the minimum values of these two sets and the maximum values for these two sets for each variable. All three ratios are expected to be equal to one if the two groups (Group 1 = not missing versus Group 2 = missing) are the same.

Chart 6: Ratios of Mean, Minimum, and Maximum for Observations with Some Missing versus Non-missing



⁸CDM Report, page 6-45

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

50. There are only eight variables out of the 26 where there isn't some serious difference between the two groups (missing and non-missing). Variables on the right side of the chart have significant differences in the maximum values, meaning the range of values for one group is truncated relative to the other. Variables on the left side of the chart have significant differences in the minimum values, meaning again that the range of values for one group is truncated relative to the other.

51. **This is the fifth key problem in Dr. Olsen's analysis.** These inconsistencies mean that the data is severely biased by missing values. Observations that are missing some data are unlike those that are not missing data. Analysis of a data set with characteristics like this is fruitless since there is no way to know what the real relationships are in the data. Since PCA relies on these relationships, the PCA conducted by Dr. Olsen is meaningless.

Methods for Treating Missing Data

52. Dr. Olsen substitutes the means for the missing data. This forces the distribution of the data to change since different variables have different numbers of missing values. It would seem that this process would reduce the variability in the data set, but in fact it may artificially reduce the variability on an individual variable. At the same time it will also inflate or deflate the correlation between two variables, and change the direction of the correlation.

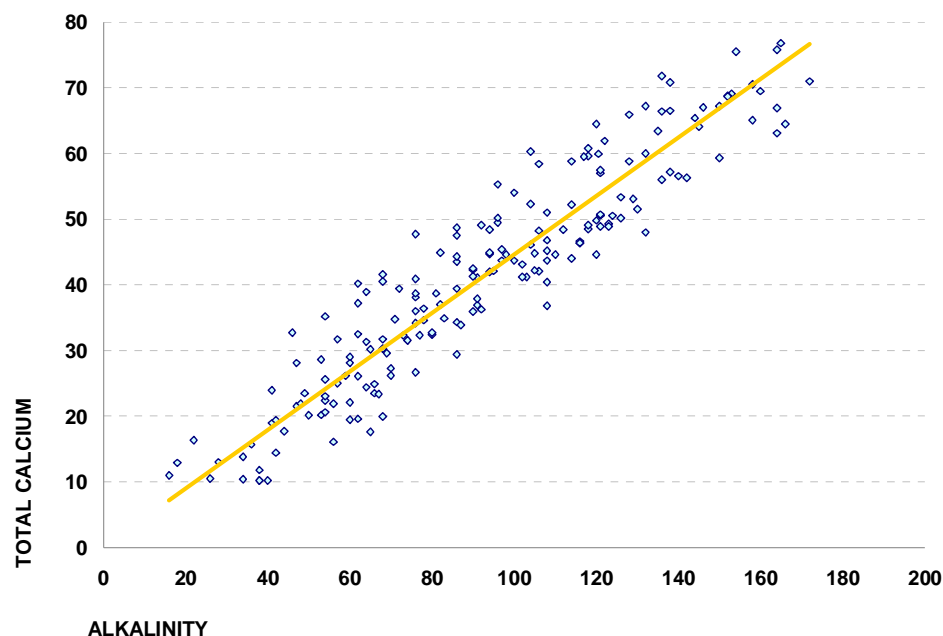
53. An example taken from Dr. Olsen's data follows. Chart 7a shows the relationship between calcium and alkalinity for observations where both values are observed.

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Chart 7a:

Complete

Observations



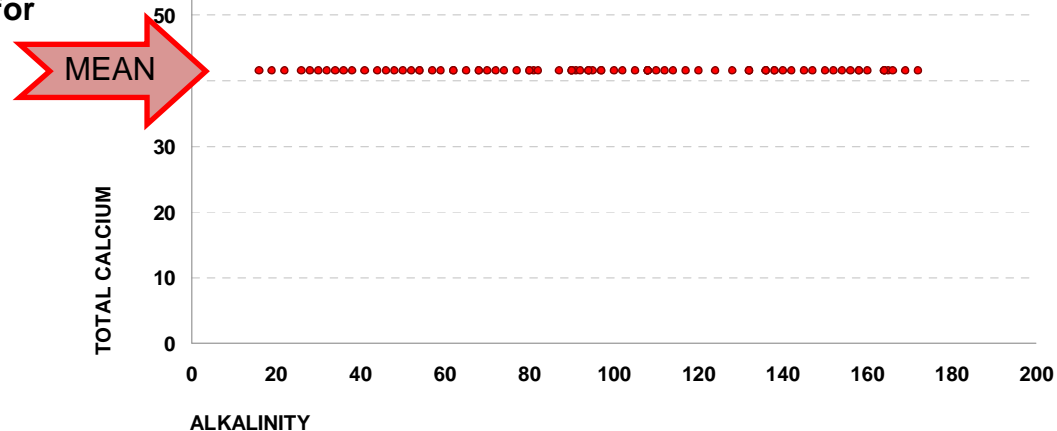
54. Dr. Olsen is missing a large number of observations on both Calcium and Alkalinity. When he is missing an observation, he substitutes the mean, regardless of what he knows about the other variable. In other words, if he is missing a value on Calcium, he plugs in the mean regardless of anything he knows about alkalinity. The same is true for alkalinity.

Chart 7b:

Missing Values

Plugged In for

Calcium



REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Chart 7c:

Missing

Values

Plugged

In for

Alkalinity

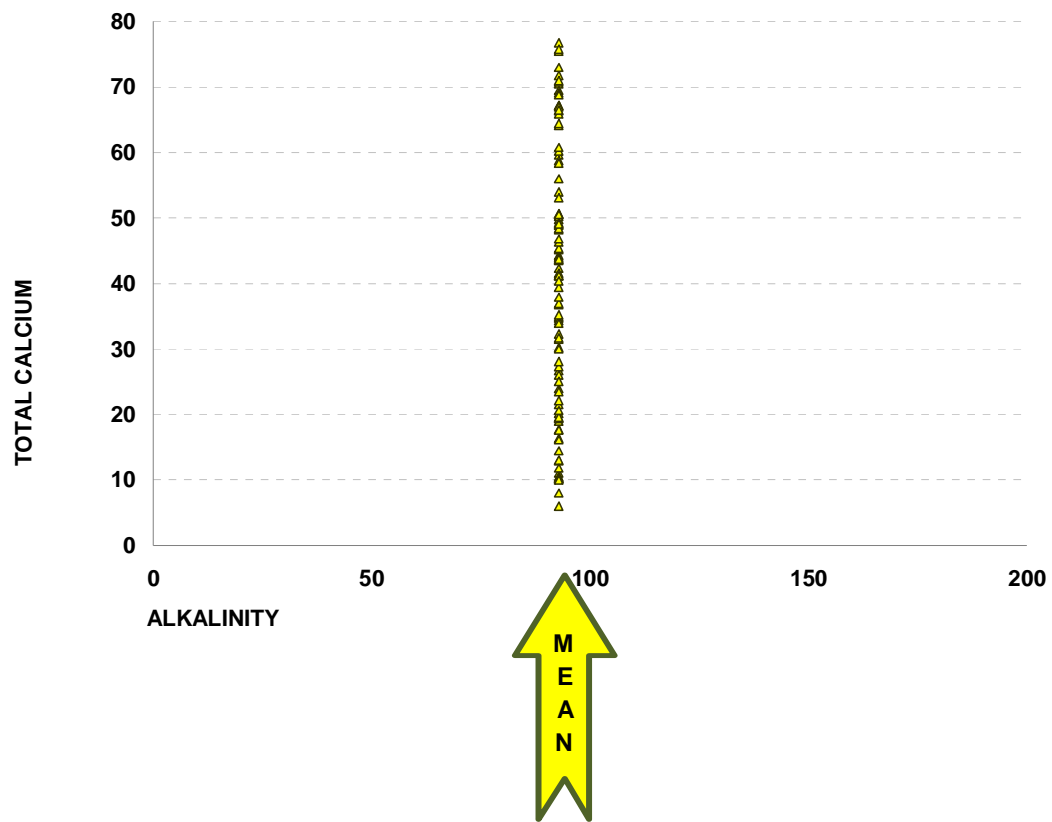
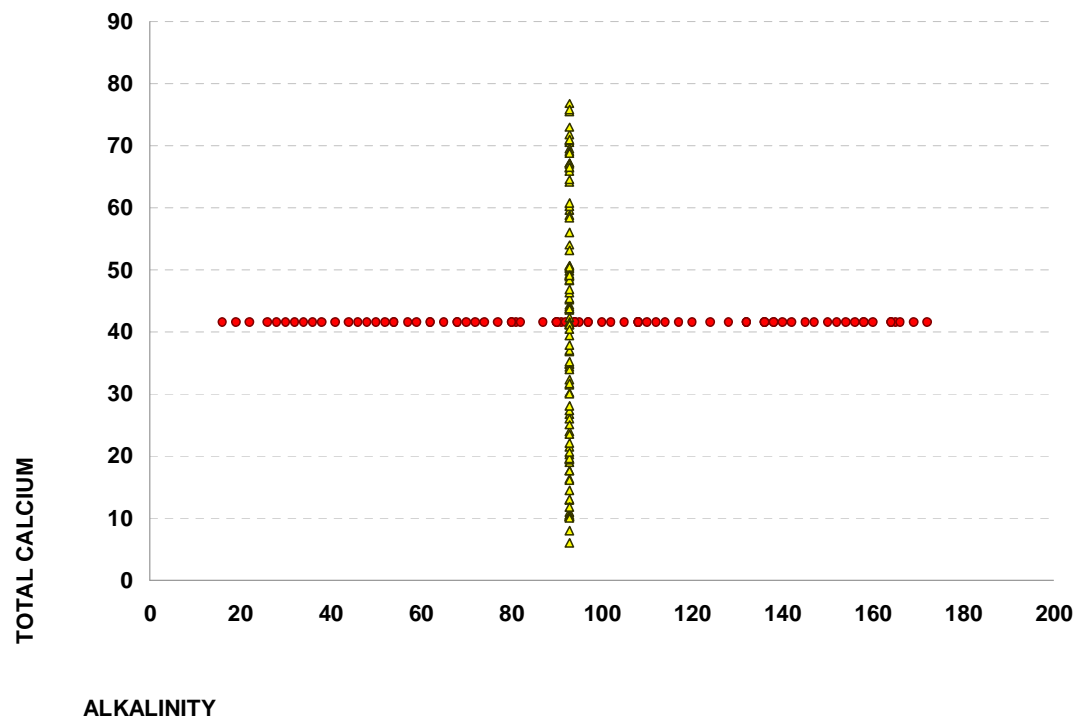


Chart 7d:

Combination

of Missing

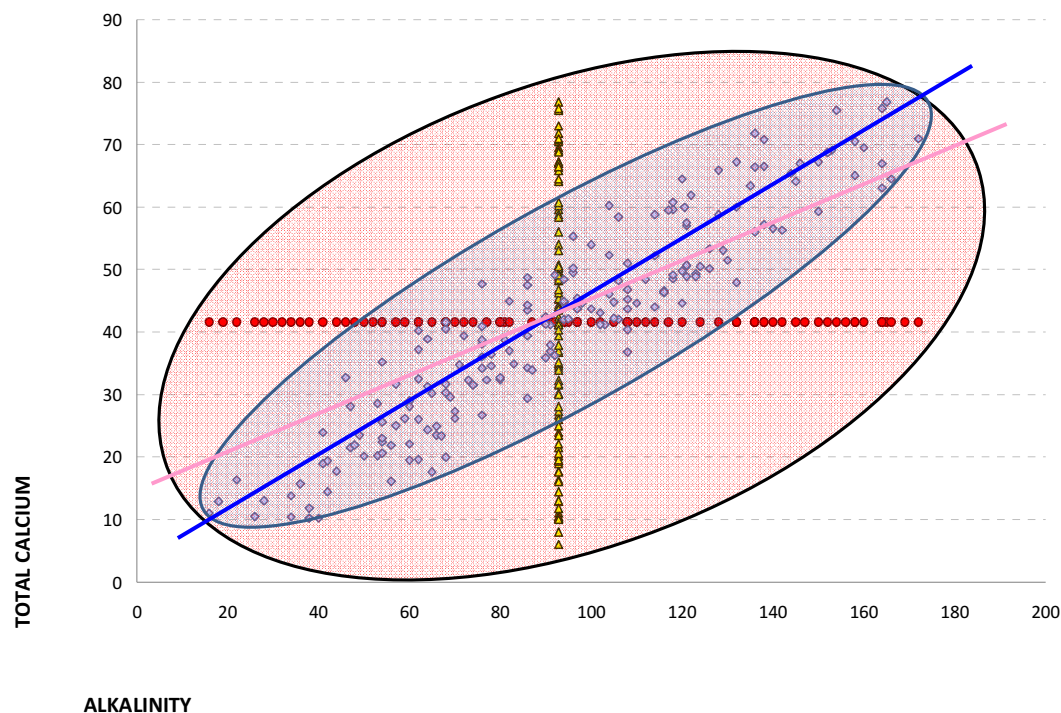
Values



REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Chart 7e: Combination of Missing Values with Known Values – the Data Set

Analyzed by Dr. Olsen



The larger pink ellipse covers what Dr. Olsen analyzed, but it is skewed from the real data and has a much greater artificial variability. The narrower blue ellipse is the original data.

55. Any line used to describe a relationship is skewed by the amount of missing data substituted and the differences in the ranges and means of each set of data (missing and nonmissing).

56. This is the sixth key problem in Dr. Olsen's analysis. His method of substituting for missing data skews the relationships in the data. At the same time, Dr. Olsen's method of substitution inflates variances, again changing the relationships being measured. These two outcomes make it impossible to measure any true relationships in the data. Dr. Olsen has hidden the true relationships by changing them with missing data substitutions designed to hide defects in his data and his calculations.

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Non-Detects

57. In the data analyzed by Dr. Olsen, he also has a number of values that are non-detects, meaning the measurement method used by the researchers cannot measure any trace measure of a chemical or organic value. Rather than treat this as a zero (not detected), Dr. Olsen substitutes the midpoint between zero and the detect limit for a chemical⁹. However, the detect limits can vary from observation to observation for each chemical. In some samples we would have a smaller non-detect than for others, such as .01 as a lower limit for some observations on Aluminum, and .001 for other lower limits. This variability in detection levels adds to the variability in the data, exacerbated by the use of logarithms. This is another method of treatment of missing data, but the impact will be discuss later in this report.

USGS vs. non-USGS observations

58. As noted previously, Dr. Olsen takes observations from the USGS¹⁰ and combines them with observations from the plaintiffs and treats them all as if they are measuring the same relationships, but he does so without testing if there is a difference between the two datasets.

59. **This is the seventh key problem in Dr. Olsen's analysis.** Ignoring the sources of the data ignores any incompatibility in the data. The table below replicates Dr. Olsen's analysis exactly for the PCA, but conducts his analysis twice – once for the USGS cases and once for the non-USGS cases. The rotated factors are presented.

⁹ CDM Report page 6-40 and page 6-47

¹⁰ CDM Report, page 5-1 and page 6-38

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Table 1: Analysis of Two Separate Parts of the Data Collected

NOT USGS						USGS					
Variables	1	2	3	4	5		1	2	3	4	5
MN	0.836	0.102	-0.035	-0.148	0.068		0.897	-0.079	-0.074	0.204	0.046
FE	0.853	0.205	-0.187	0.124	-0.169		0.864	-0.172	-0.013	0.210	0.267
AL	0.787	0.217	-0.263	0.170	-0.209		0.846	-0.162	0.072	0.208	0.284
NI	0.762	0.135	0.224	0.236	-0.049		0.774	0.343	0.103	0.150	-0.222
AS	0.745	0.038	0.106	-0.018	0.013		0.767	0.133	0.303	0.190	-0.342
BA	0.590	-0.032	-0.333	0.016	0.314		0.701	0.320	0.251	0.142	-0.307
CU	0.698	0.308	0.152	0.337	-0.175		0.784	-0.057	0.129	0.085	-0.124
ZN	0.688	0.081	0.093	0.271	0.033		0.881	-0.069	0.070	0.111	0.013
TOC	0.607	0.439	0.306	0.167	-0.223		0.726	0.033	0.028	0.379	0.044
P_SOL_REAC	0.253	0.061	0.318	0.814	-0.079		0.056	0.388	0.861	0.125	0.038
P_TD	0.304	0.089	0.377	0.786	-0.119		0.339	0.307	0.832	0.166	-0.148
P	0.558	0.141	0.283	0.684	-0.126		0.577	0.185	0.702	0.242	-0.139
NO2_NO3	-0.144	-0.039	-0.086	0.734	0.233		-0.229	0.202	0.704	-0.207	0.432
FECAL	0.095	0.954	-0.014	0.048	-0.062		0.316	-0.101	0.097	0.848	0.145
COLIFORMS	0.147	0.913	-0.013	0.039	-0.086		0.312	-0.131	0.180	0.603	0.410
ENTERO	0.130	0.886	-0.035	0.033	-0.090		0.398	-0.195	0.069	0.753	0.273
ECOLI	0.121	0.814	-0.032	0.018	0.026		0.186	0.008	-0.010	0.874	-0.101
CA	-0.133	-0.168	0.183	0.000	0.882		-0.272	0.815	-0.120	-0.107	-0.100
ALKALINITY	-0.100	-0.077	0.216	-0.093	0.835		-0.343	0.626	-0.269	-0.067	-0.075
TDS	0.282	0.003	0.219	0.324	0.476		-0.104	0.871	0.176	0.007	-0.154
SO4	0.109	0.018	0.802	0.110	0.113		0.076	0.864	0.367	-0.083	0.088
NA	-0.223	-0.087	0.837	0.190	0.211		0.063	0.914	0.309	-0.050	-0.004
CL	-0.154	-0.062	0.753	0.217	0.310		-0.021	0.910	0.285	-0.077	0.008
MG	0.492	0.091	0.618	0.077	0.153		0.317	0.757	0.150	-0.077	0.069
K	0.519	0.139	0.537	0.466	-0.097		0.454	0.748	0.412	0.050	-0.046
TKN	0.271	0.322	0.220	0.013	-0.122		0.072	0.028	-0.008	-0.369	-0.717

Variables in yellow are the six variables that were measured differently by the USGS and the plaintiffs.

60. Results change significantly for variables that differ in measurement between USGS and the plaintiffs' collection. Note that for the three phosphorus measures, they are only in the fourth principal component for the non-USGS data, but in the third principal component for the USGS data. This means the supposed importance of the phosphorus measures is lower for one data set than another – this shouldn't happen if the two datasets are equivalent.

REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

61. For the plaintiffs' measurement of **total dissolved solids** (TDS), TDS doesn't surface on ANY of the principal components, meaning it is not important to any of the factors measured. However, for the USGS measurement, it is a key component of the second principal component. Similarly, in the measurements by the plaintiffs, calcium and alkalinity contribute a completely separate factor, uncorrelated with a principal component that includes sulfate, sodium, chlorine, and magnesium. For the **USGS**, there is a single component that combines calcium, **alkalinity, total dissolved solids** with the factor measured by the plaintiffs. **Again, this shouldn't happen if the two sets of data are equivalent in the way they measure constituent elements.** Finally, **TKN** doesn't carry any weight in defining principal components in the data from the plaintiffs, whereas in the USGS data it is so important that defines it's own principal component, again separate from the remainder of the components.

Use of Logarithms

62. Dr. Olsen converts all of his observations by taking logarithms of values before conducting the PCA¹¹. Use of logarithms in statistical analysis is common in a number of fields and is usually done for one of three reasons. One reason is to stabilize the variability of the data so that the data more closely follows a particular statistical distribution. As Dr. Olsen didn't conduct any statistical tests, this can't be the reason.

63. The second reason is to transform data with exponential relationships to data with linear relationships, as methods for analyzing linear data are much easier to employ. This may be the case here, but there are costs for doing so and there is no discussion regarding why data in the water samples would have multiple exponential relationships.

¹¹ CDM Report, page 6-46

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

64. The third reason is to reduce the natural variability of data and pull in outlying observations. When this is done, it typically disguises problems in data collection or unusual observations that should have been separately analyzed. This is certainly the case with this data.

65. The problem with the use of logarithms is that it significantly reduces the variability of the observed data and changes the correlation between the observations. In the extreme, two variables that have no linear relationship (correlation of zero) can have a perfect correlation when one takes logarithms. This means that a PCA of data where logarithms are taken will result in a completely different outcome than a PCA of the original data.

66. Dr. Olsen doesn't explain why he takes logarithms, he simply does so. There is no examination of whether correlations measured on the logarithmic scale also exist in the real world. Dr. Olsen doesn't consider the interpretation of a principal component once he has conducted an analysis.

67. As described in an earlier section, each principal component is a weighted sum of the variables in the analysis. A principal component is written as:

$$\text{Principal Component} = c_1V_1 + c_2V_2 + c_3V_3 + \cdots + c_{26}V_{26}$$

where the coefficients c_j are related to those presented in the table above in the USGS \ non-USGS analysis or any of the other PCA analysis. However, this would be true for those cases where the variables V_j are in their original form. Now suppose we have a principal component that is on the logged values.

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

68. In Dr. Olsen's analysis, we have:

$$\text{Principal Component} = c_1 \text{Log}(V_1) + c_2 \text{Log}(V_2) + c_3 \text{Log}(V_3) + \dots + c_{26} \text{Log}(V_{26})$$

69. Using a simple algebraic result, we transform this equation into one involving the original data

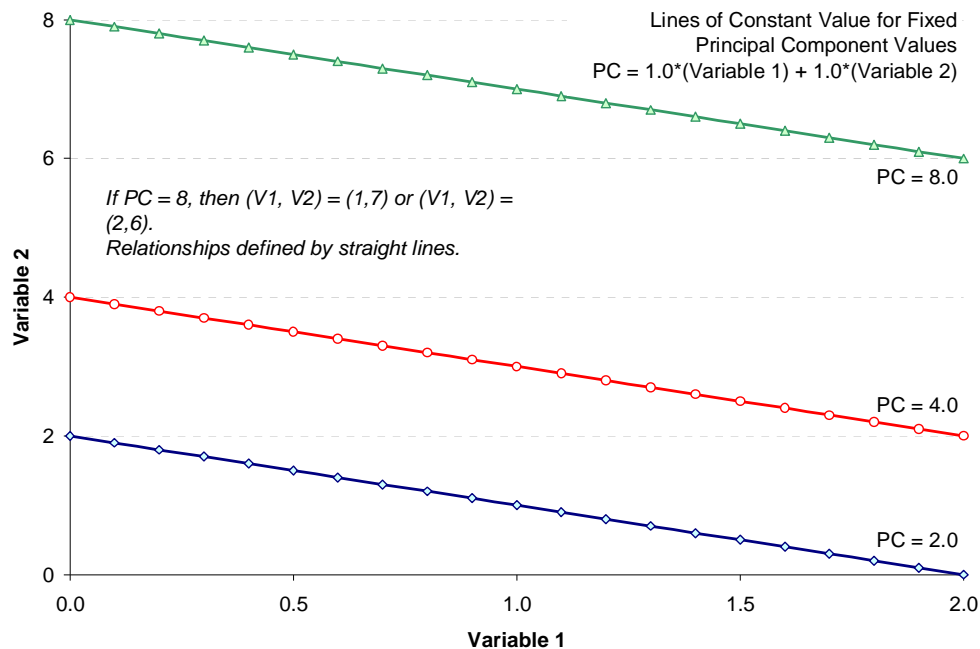
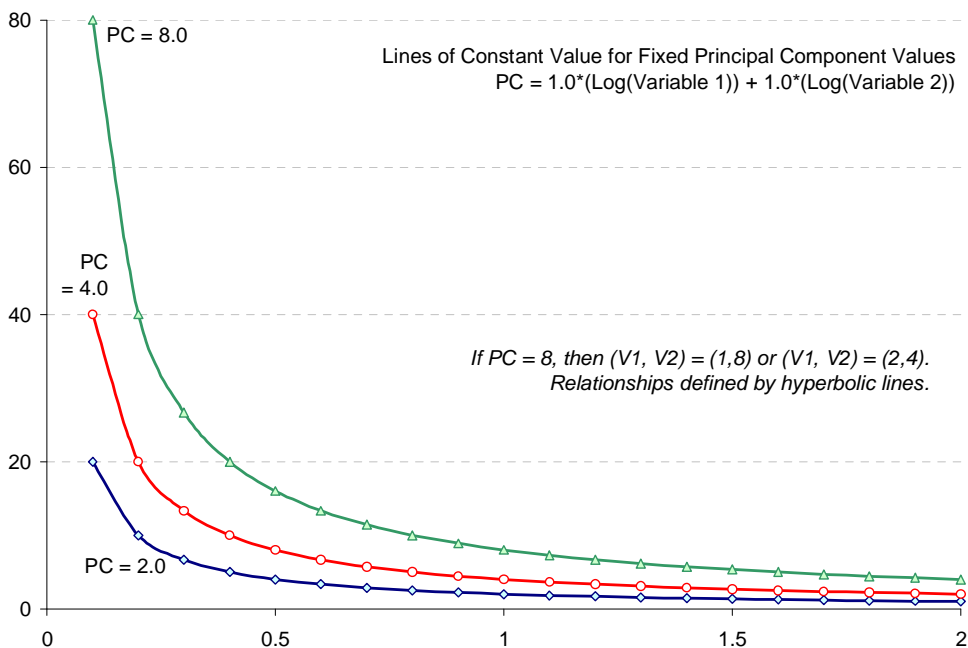
$$\begin{aligned} \text{Principal Component} &= c_1 \text{Log}(V_1) + c_2 \text{Log}(V_2) + c_3 \text{Log}(V_3) + \dots + c_{26} \text{Log}(V_{26}) \\ &= \text{Log} \left[(V_1)^{c_1} * (V_2)^{c_2} * (V_3)^{c_3} \dots (V_{26})^{c_{26}} \right] \end{aligned}$$

70. With the logarithms, the principal component is NOT a sum of the variables. The principal component is the **product** of the variables, each raised to some factor that weights it. Because it is a product, this means that any findings do not relate back to any findings in the real world in the ways that Dr. Olsen describes. Results from Dr. Olsen's analysis are multiplicative, not additive. Dr. Olsen mistakenly ignores this outcome in his transformations.

71. Charts 8a and 8b show the contrast in the relationships. If Dr. Olsen had not used the logarithms, his relationships between variables would be straight lines. A principal component would represent the sum of values (like a measure of iron plus a measure of aluminum plus a measure of copper within one sample). But Dr. Olsen did use logarithms, which forces all the relationships to be curved. Worse, for a set value in the principal component analysis, the outcome is either a very large amount of variable one combined with a very small amount of variable two (lots of iron, very little copper) or a large amount of variable two combined with a very small amount variable one (very little iron and lots of copper). For the most part, using logarithms, a fixed value for a principal component represents extremes of one variable or another, but not of both variables, completely undercutting his argument that his results represent a "signature".

REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Chart 8a: Linear Relationships From Use of Actual Variables in Principal Components**Chart 8b: Curved Relationships Implied by Logarithmic Transforms of Variables Used in Principal Components**

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

72. There are other problems with the use of logarithms. One is that, in trying to fit a relationship between two variables, observations receive different weight for their contribution to the relationship if log values are used compared to when the original values are used. This means that there are values that will have a strong effect on the outcomes when used as an actual value. The same values will not have an effect on the outcomes if logged, while other observations will have a stronger effect than would have happened with the original data. Because of this, use of logarithms has to be done with great caution since the interpretation of the value of the inputs differs greatly. A particular example of this is found in the non-detects.

73. As noted before, the **non-detects** have their importance greatly heightened in the analysis. The logarithm of a number is the exponent of the number represented as raised to the power of ten. The table below demonstrates what the values are:

Number	0.000001	0.0001	0.01	1	100	10000	1000000
Equals	10^{-6}	10^{-4}	10^{-2}	10^0	10^2	10^4	10^6
Logarithm	-6	-4	-2	0	2	4	6

74. A non-detect of .01 versus a non-detect of .001 might not seem like much a difference, but in the log scale this can be the difference between -2 and -3. If the variable being measured typically has values in the range of 10 to 100 milliliters, the value being analyzed on the log scale is somewhere in the range of 1 to 2. A change in the non-detect value of -2 to -3 (merely because of very minor differences in the test) will have huge effects on the outcome.

REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

75. This is the eighth key problem in Dr. Olsen's analysis. He doesn't perform any sensitivity analyses to determine if the non-detect limits affect his outcomes. If most of the values for a logged variable range from 1 to 2 and then an arbitrary value of -2 or -3 is thrown into the analysis, he has created outliers that leverage the relationship. Two variables with a straight line relationship, both measured on a scale of 1 to 2, will be greatly impacted by values thrown in at the far end of the scale. Furthermore, why chose the midpoint between zero and the non-detect value as the substitute value? Why not another value, closer to zero or closer to the non-detect value? Since the log transformation has such power in moving the end of the relationship, the impact of this choice should also have been measured.

The Number of Principal Components and Rotations

76. The final analytical issue for discussion is the number of principal components that came out of the analyses and their meaning. Dr. Olsen conducted the PCAs as described above, but he only retains the first two principal components. He throws away significant results that may explain patterns not found in the first two components¹². These later components are the ones that may be most useful in explaining specific results.

77. Further, he arbitrarily reports on non-rotated factors at times and ignores the rotated outcomes. The problem with doing this is that a non-rotated factor is measuring a distance in a way that cannot be interpreted (see the earlier description of this problem). Dr. Olsen's data show the problems with both of these actions. The following table presents the outputs from Systat (the program he used) using the data from his datasheets.

¹² "These variances indicate that PC1 and PC2 are by far the most important of the five together explaining 56.2% of the total variance, relative to PCs 3, 4, and 5 (17.8%)", CDM Report, page 6-51

REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Table 2: The Five Principal Components from Dr. Olsen's Analysis

CU_T	0.851	-0.032	-0.077	-0.070	-0.047	0.161
P_T	0.812	0.341	-0.057	-0.313	0.142	0.033
TOC	0.812	-0.040	0.110	0.005	-0.175	-0.044
NI_T	0.801	0.106	-0.216	0.078	-0.089	-0.073
FE_T	0.797	-0.332	-0.308	0.021	0.083	-0.203
AL_T	0.765	-0.367	-0.277	-0.043	0.129	-0.195
K_T	0.743	0.473	0.010	-0.135	-0.138	0.020
ZN_T	0.721	-0.075	-0.175	0.130	-0.021	0.248
AS_T	0.672	-0.063	-0.304	0.214	-0.135	0.118
MN_T	0.658	-0.206	-0.385	0.304	-0.029	-0.250
P_TD	0.637	0.511	0.072	-0.423	0.162	0.099
MG_T	0.575	0.422	-0.015	0.259	-0.257	-0.025
P_SOL_REAC	0.559	0.526	0.076	-0.438	0.240	0.115
NA_T	-0.003	0.838	0.259	0.064	-0.223	-0.207
CL	0.036	0.816	0.231	0.132	-0.142	-0.157
SO4	0.243	0.696	0.177	0.102	-0.298	-0.313
TDS	0.302	0.474	-0.092	0.247	0.269	0.131
FECAL	0.554	-0.380	0.651	0.170	0.118	-0.037
COLIFORMS	0.556	-0.370	0.603	0.134	0.103	-0.041
ENTERO	0.552	-0.406	0.578	0.148	0.116	-0.093
ECOLI	0.481	-0.321	0.547	0.231	0.106	0.067
BA_T	0.381	-0.108	-0.460	0.287	0.292	-0.139
CA_T	-0.252	0.538	-0.053	0.609	0.351	0.070
ALKALINITY	-0.227	0.478	0.000	0.649	0.240	0.206
NO2_NO3	0.044	0.406	0.070	-0.320	0.578	0.041
TKN	0.347	-0.044	0.020	0.092	-0.355	0.689

* Factor loadings above 0.6 are in red, factor loadings above 0.45 are in blue if there are no other factors in red for the same variable.

78. Using Dr. Olsen's methods, we would throw away the principal component that has bacteria (fecal, coliforms, entero, and e-coli). But the other experts for the plaintiffs claim this to be the most important data for analysis of chicken waste. This inconsistent treatment of key information raises the question about what is significant and whether there is any consistent treatment of the data produced by Dr. Olsen.

REPRODUCING THE SW3 DATA RECORDS AND VALUES

79. Dr. Olsen used a program called EDA_Analyzer to capture the data from the main database and loaded the data into an Excel worksheet referred to as SW3. It appears that he substitutes means for the missing values (see the earlier discussion in this report on this point). Dr. Olsen then takes logarithms of the SW3 values before using Systat to calculate PCA loadings (coefficients). The results of the Systat loading coefficients are transferred to an Excel sheet and he calculates the PCA values presented in Appendix F of the CDM report¹³.

80. We attempted to reproduce the values in the SW3 Excel sheet and the PCA values in Appendix F of the CDM report. All of the records from the master database with “SW:S” in the sample groups were downloaded into an Excel file. This download produced an Excel sheet with all of the surface water data. We had to make some changes in sample group identifications to match the EDA_Sample IDs found in Appendix F of the CDM report. Some of the changes were to add USGS to the sample group IDs that only had numbers. There were other changes made to the sample group IDs that involved removing blank spaces and changing noncapital letters to capital letters. This work was required to be able to finally link the values reported by Dr. Olsen in his written report to the same values in Dr. Olsen’s data – there was little correspondence between values in the written report and the database and it required a significant effort to be able to link which data records Dr. Olsen selected from all of those available. I revisit this topic later as there seems to be little consistency in choices made for the data ultimately included in the analysis. We picked the appropriate measurement unit for values that were measured in UG/L units and we used the P0065 measurement values for the USGS variables TKN, TDS, SO4, P_TD, P_T and P_SOL_REAC.

¹³ CDM Report, page 6-53

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

81. In the CDM report, the names of the variables are used in Dr. Olsen's descriptions. The database also has the names of the variables, but a "ParamID number" is also associated with each variables. If the CDM report in Dr. Olsen's tables presented the ParamID number along with the name of the variable, then it would be clear which variable is being discussed. This is an issue because it is nearly impossible to scale down from the 315 variables in the Access database to the 26 in the Excel database. There is no documentation as to exactly which variables were extracted by Dr. Olsen or his subordinates, and only through diligent detective work was it possible to work backwards to discover which 26 variables were selected. As will be seen in a later section, there is no standard data selection procedure that would indicate how Dr. Olsen got from the 315 variables in his full database to the final 26 variables he selected. In fact, given how much information is missing in the database, the final set of 26 variables used is counterintuitive.

82. There are 26 variables in the final SW3 Excel spreadsheet analyzed by Dr. Olsen¹⁴. Each variable has a parameter key in the database table RefParm that indicates the name of the variable. Appendix F in the CDM report has a listing of the 573 samples (EDA_Sample) used in the PCA runs for the SW3 data. The EDA_Sample IDs are produced from combining several sample keys and sample groups into one sample group, or as referred to in the CDM report, an EDA_Sample. Below is a small example of what occurs when the data is downloaded from the database. There are usually several samples for each sample group.

¹⁴ CDM Report, page 6-45

REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

		Variable IDs					
sampleky	sample group	4	8	39	42	58	59
105025	BS-08:8/23/2005:SW:S:-:-						
105178	BS-08:8/23/2005:SW:S:-:-	98	6.53				
105179	BS-08:8/23/2005:SW:S:-:-				0.134	0.5	0.014
106374	BS-08:8/23/2005:SW:S:-:-			68			
106848	BS-08:8/23/2005:SW:S:-:-						
105189	BS-117:9/14/2005:SW:S:-:-	132	10.18				
105190	BS-117:9/14/2005:SW:S:-:-				0.42	3.23	0.038
106079	BS-117:9/14/2005:SW:S:-:-						
106175	BS-117:9/14/2005:SW:S:-:-			13000			
106849	BS-117:9/14/2005:SW:S:-:-						

83. The rows of data for any given sample group are collapsed into only a single row. The rows were collapsed by moving values into missing areas in the first row of a given sample group. A sample group that has more than one value for a variable is averaged¹⁵. Below are the collapsed rows for the example given above.

		Variable IDs					
sampleky	sample group	4	8	39	42	58	59
106848	BS-08:8/23/2005:SW:S:-:-	98	6.53	68	0.134	0.5	0.014
106849	BS-117:9/14/2005:SW:S:-:-	132	10.18	13000	0.42	3.23	0.038

84. The sample keys (the first column) are not indicated in the SW3 Excel sheet because all of the sample keys have been collapsed into individual samples. We were able to match all of the EDA_Samples in Appendix F with the collapsed sample groups **but not all of the values**. Dr. Olsen's SW3 Excel sheet has 573 rows with 26 variables; therefore his sheet has 14,898 values. His SW3 data has 915 missing values. The SW3 Excel sheet we produced has 573 rows and the same 26 variables, but the composition of the entries is very different.

¹⁵ CDM Report, page 6-47

REBUTTAL REPORT

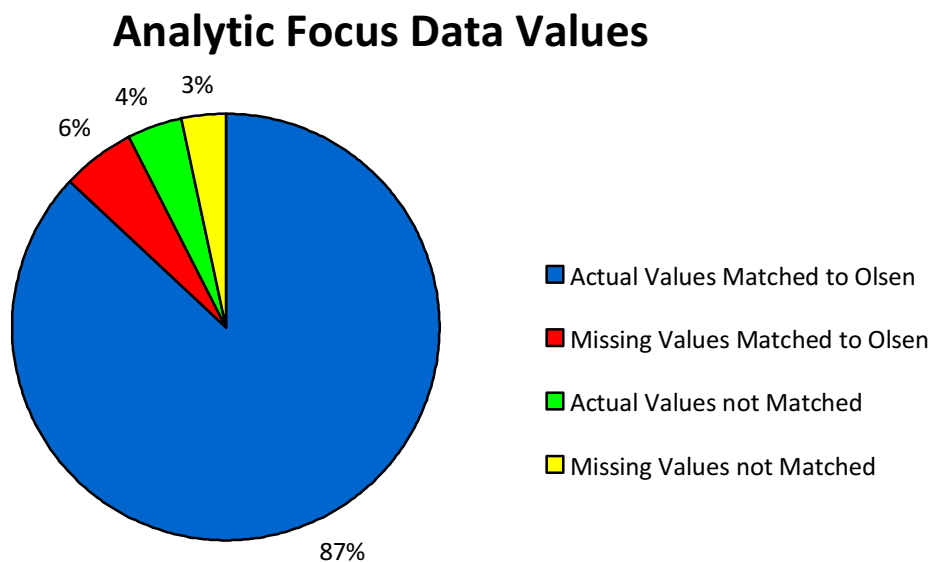
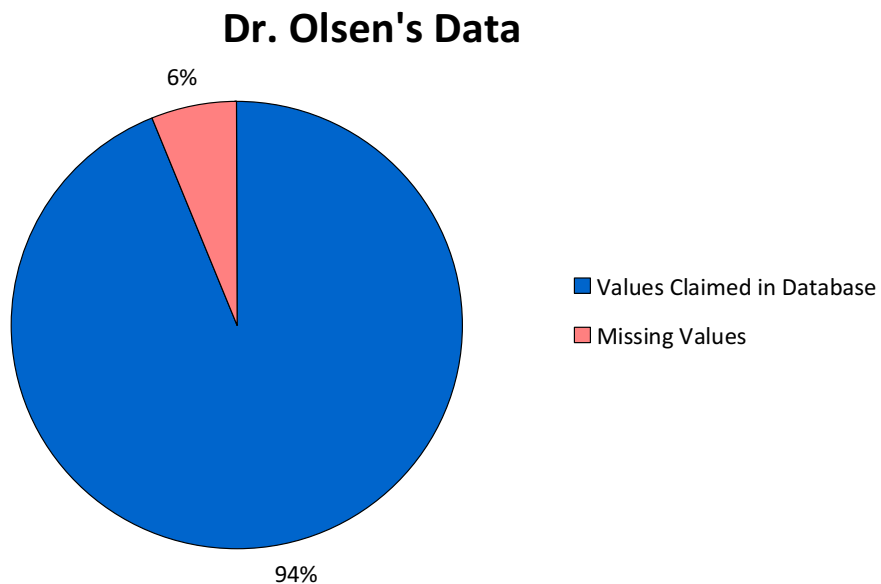
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Dr. Olsen's SW3 data	13,983 values + 915 missing values = 14,898 values
<hr/>	
Analytic Focus	12,933 matched data values (Agreement with Olsen)
+	849 matched missing values (of the 915) (Agreement with Olsen)
+	499 missing values (Database is missing, but Olsen has data)
+	66 data values exist (Database has data, but Olsen has missing)
+	551 non-matched values (Database and Olsen's Excel Files differ)
=	14, 898 values

85. To summarize, of the 915 missing values that Dr. Olsen had, we found only 849 missing values – the remaining 66 were decreed by Dr. Olsen to be missing when they in fact had data. In addition, there are 499 additional values that were missing data in the Access database, but which suddenly have data in Dr. Olsen's analysis file. Finally, there are 551 values in the dataset where the value in the Excel file used for analysis differed from the original values in the Access database. In total, there are over 1,000 cells in Dr. Olsen's analysis database that do not correspond to the original data. This is about 7.5% of the total data that is in error or changed in some manner. This calls into question any quality of any analysis or data used by Dr. Olsen. Additionally, the 1,116 cells that have discrepancies are only one part of the problem. There is also a significant amount of data thrown away or ignored for no discernable reason.

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

86. These outcomes are summarized in the next two charts.



REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

87. Dr. Olsen calculates his PCA scores in Appendix F of the CDM report in an Excel sheet ("To calculate a PC score for each individual sample, the PC coefficient is multiplied by the standardized parameter concentration. This is performed for all parameters (variables) (*sic*¹⁶) in a particular PCA run. The product values for all 25 (*sic*¹⁷) parameters are summed to yield one PC score for each sample for each PC. Hence, a particular sample will have both a PC1 and a PC2 score").¹⁸ We reproduced Dr. Olsen's PCA scores in the following manner.

88. Start with the original SW3 data for the 26 variables. Missing values are replaced with the means of the variables before taking the logarithms. Compute z-transformations (subtract the mean of a variable, divide by it's standard deviation) on these original variables. Multiply the SW3 z-transformed variables by the first two sets of coefficients produced from Dr. Olsen's PCA on the SW3 log base ten data, ignoring the remaining sets of PCA coefficients. This produces two variables with 573 observations each. The 573 observations are the EDA_Samples (S1, ..., S573). The two variables are PC1 and PC2.

89. To calculate PC1 for the first EDA_Sample "S1", find the minimum value of the PC1 column, take the absolute value of the minimum value, add 1 to this value, then add the value of the first EDA_Sample. This method does not correspond to any standard PCA methodology.

¹⁶ Dr. Olsen throughout his report confuses the terms parameter and variable. In this sentence he uses one to explain the other. From context, it seems that Dr. Olsen means variable when he says parameter. A parameter is a single value that describes a characteristic of a population, like an arithmetic mean or a variance. A variable is a theoretical construct used to denote a value that can change according to the sample being observed. These are not interchangeable terms.

¹⁷ There are 26 variables in Dr. Olsen's analysis, not 25.

¹⁸ Dr. Olsen's calculations are described on page 6-53 in the CDM report

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

90. The calculations above in the previous two paragraphs used to duplicate the PCA values do not match the description of how to calculate PCA values given in the CDM report. The CDM report does not describe taking the absolute minimum of a column as a part of the calculation. Using this procedure, we were able to exactly replicate the scores used by Dr. Olsen.

91. In this process, Dr. Olsen commits an error so basic and so egregious that it completely invalidates every result and conclusion he offers. He runs the PCA on the logarithmic scores, but he ignores the SYSTAT program's calculations and instead applies the PCA coefficients to the original data without taking the logarithms.

92. Just to be clear, I will repeat the steps taken by Dr. Olsen for analysis:

- a. Take original data in the dataset with no missing data (26 variables)
- b. Take the logarithm base 10 of the values in the original data (26 new variables)
- c. Compute a transformation on the log values as (26 new variables again):

$$\text{New data value} = \frac{\text{Logged Data Value} - \text{Mean of Logged Data Values}}{\text{Standard Deviation of Logged Data Values}}$$

- d. Use the variables created in step c. to run the PCA in Systat
- e. Save the coefficients from Systat to apply to a different set of input data to compute scores for PC1 and PC2
- f. Compute scores for PC1 and PC2 outside of Systat using coefficients from step e. applied to data that skips step b. above.
- g. Translate scores for PC1 and PC2 from scale in step f. so that it appears there are no negative scores.

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

93. **Systat readily computes the scores that Dr. Olsen wants.** However, Dr. Olsen ignores this and uses the coefficients from the analysis of the logarithmic data, but applies these coefficients to the original data without logarithms. What should have happened is that the coefficients should have been applied to the logged data to compute the scores.

94. To give a sense of the order of magnitude of this error, consider the problem of sending a rocket to Mars. The distance of Earth to Mars is a maximum of 250 million miles, which occurs when the planets are on the opposite sides of the Sun. On the log base 10 scale, the one used by Dr. Olsen, 250 million miles translates to 8.398. Remember that the logarithm computes the power of 10 needed to find the number of interest. So $10^{8.398} = 250,000,000$. Now compute fuel requirements on a distance of 8.4 miles (to be generous) and send the rocket off to Mars. The rocket would peak at just above 8 miles (not the 250 million needed) and then fall to Earth since it wouldn't even clear the atmosphere. This is the calculation that Dr. Olsen has done.

95. Dr. Olsen computes all of his coefficients on the logged data and then applies the coefficients to the original data ignoring the key transformation he has made. This should have been glaringly obvious when Dr. Olsen plotted his output for PC1 and PC2. The components computed are uncorrelated with one another – this is the entire basis for the computation of principal components, namely that each one is forced to be uncorrelated with all other components. The rotation methods Dr. Olsen uses enforce this – they force the rotated results to be uncorrelated, so whether one looks at rotated solutions or the unrotated solutions, they must be uncorrelated. The correlation between Olsen's PC1 and PC2 is $R = 0.31$ when it should be identically zero.

96. The correlation between Systat's PC1 and PC2 is $R = 0.0000000$, just as it should be.

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

97. This should be immediately obvious to any observer who knows anything about PCA.

Charts 9 and 10 on the next page present Dr. Olsen's score plots using his incorrect calculation and the correct score plots using the information from Systat, Dr. Olsen's program of choice.

98. Every conclusion that Dr. Olsen draws about his results that involve the use of the scores is wrong and meaningless. Computing the values that he did where there is confusion between the scales used means that Dr. Olsen had no idea what he was looking at and drew completely erroneous conclusions based on a mistake that he or his subordinates made. A quick check of the correlations between the scores would have immediately shown this error for what it was.

99. Finally, since the SYSTAT values are still on the logarithmic scale, the proper interpretation of the values would be on a real-world scale. This is easily done by computing the inverse logarithm of the Systat scores (raising 10 to the power of the score). This result is shown in Chart 11 below. When one examines this chart, one sees that there are a few extreme values charted on this plot – these result because Dr. Olsen didn't do quality control on the outliers in his data and so extremes result that are meaningless. On the proper scale, results appear on either PC1 or PC2, and there are four samples that result in extreme values in the center of the chart that are most likely due to quality control lapses.

100. Dr. Olsen's analysis, his charts, and his conclusions should be dismissed as erroneous and misleading.

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Chart 9: Olsen's PCA Score Plot

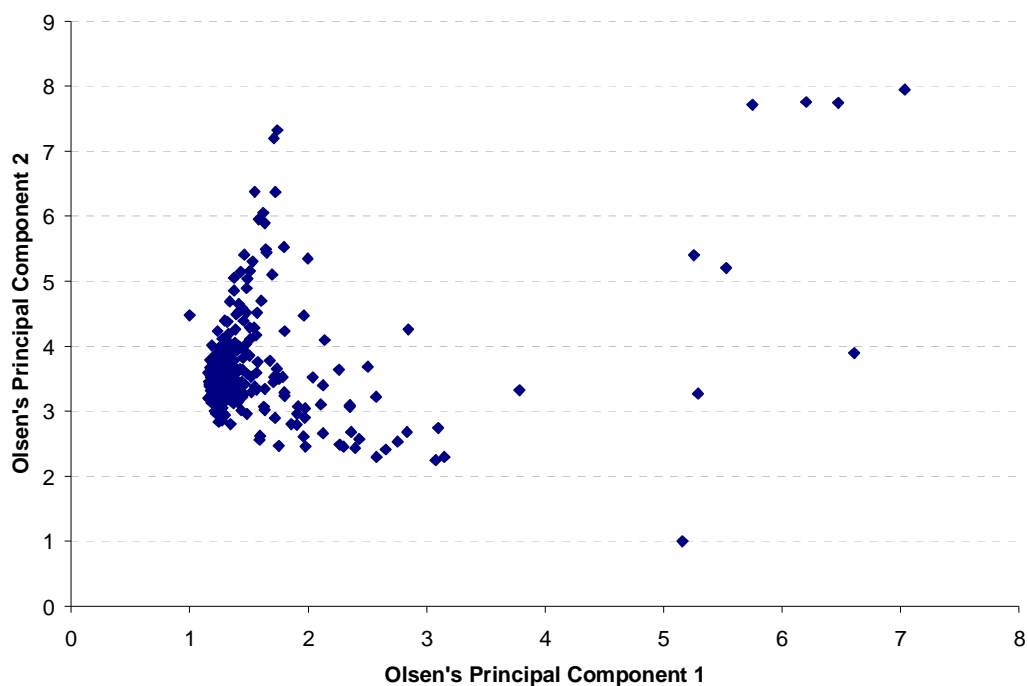
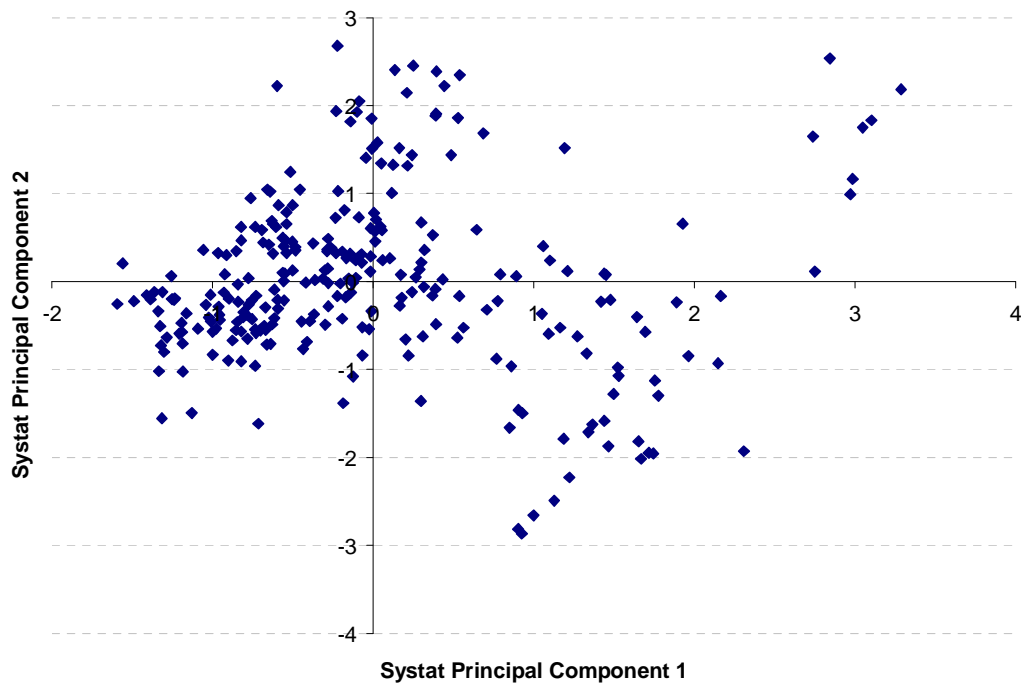
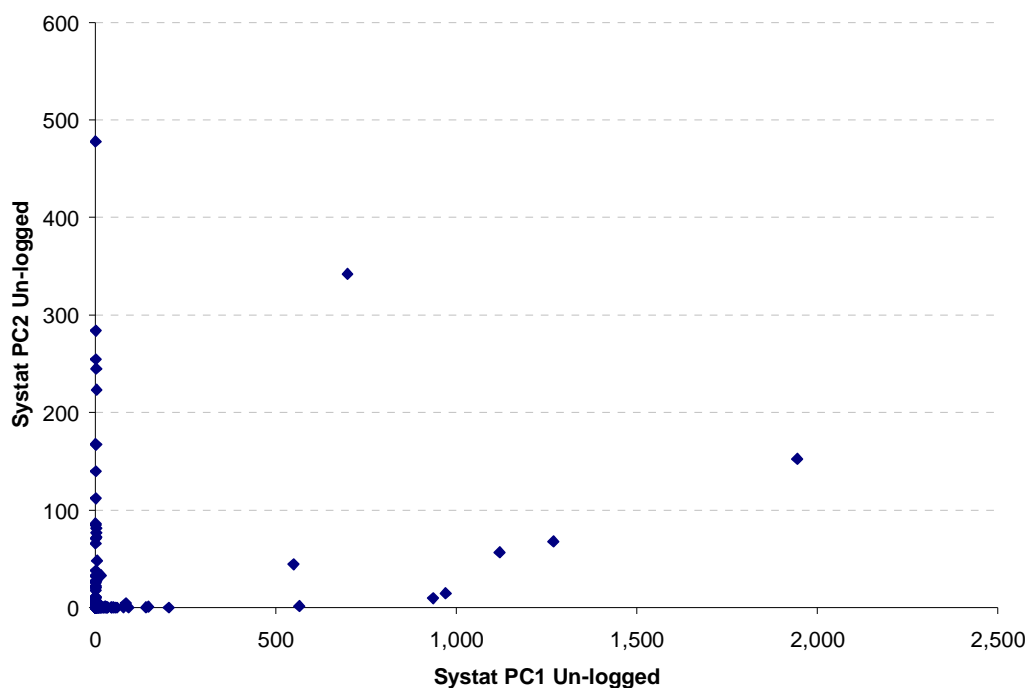


Chart 10: Systat PCA Score Plot for Principal Components 1 and 2



REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Chart 11: Systat Principal Components Converted from Logarithmic Scale to Real World Scale



Revisiting Missing Data: A File With 419 Samples And 56 Variables Without Missing Values

101. We were able to use the original 315 variables found in Dr. Olsen's original database and create an Excel sheet with 419 samples and 56 variables with no missing values.

Remember that Dr. Olsen had only 267 samples with 26 variables. In our recreation of the Excel sheet, the variables were selected based only on percentage of observations available.

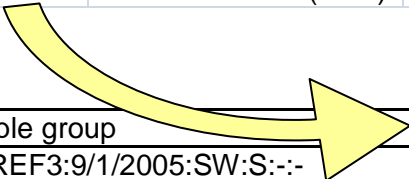
102. A query run on the database that downloads all of the records with "SW:S", which is all of the surface water data produces 66,260 rows of data. These are ported into an Excel sheet. A small example from this Excel sheet is presented below. This Excel sheet is then transformed into an intermediate Excel sheet that looks like the second example table below. This

REBUTTAL REPORT
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

intermediate Excel sheet has 6,564 rows of sample data with sample group and associated variable values.

103.

Sampleky	Paramky	ParamID	Value	SampleGrp
105152	4	Alkalinity (as CaCO3)	182	BS-REF3:9/1/2005:SW:S:-:-
105152	8	Chloride	12.44	BS-REF3:9/1/2005:SW:S:-:-
105153	42	Nitrite + Nitrate (as N)	0.368	BS-REF3:9/1/2005:SW:S:-:-



		Parameter Key			
sampleky	sample group	4	8	39	42
105152	BS-REF3:9/1/2005:SW:S:-:-	182	12.44		
105153	BS-REF3:9/1/2005:SW:S:-:-				0.368
106054	BS-REF3:9/1/2005:SW:S:-:-				
106172	BS-REF3:9/1/2005:SW:S:-:-			22	
106395	BS-REF3:9/1/2005:SW:S:-:-			42	
106871	BS-REF3:9/1/2005:SW:S:-:-				
102151	EOF07:5/15/2005:SW:S:-:-	402	30	340	0.277
106276	EOF07:5/15/2005:SW:S:-:-				
102232	EOF07:5/23/2005:SW:S:-:-				1.397
102233	EOF07:5/23/2005:SW:S:-:-				
104737	EOF07:5/23/2005:SW:S:-:-			3000	
106277	EOF07:5/23/2005:SW:S:-:-				

Paramky 39 doesn't appear until much later in the data, thus it's absence from the first table.

104. Data from the database usually has several samples for each sample group (six samples for the BS-REF3:9/1/2005:SW:S:-:- sample group in above example). The rows of data for a given sample group are collapsed into a single row. The rows are collapsed by moving values into missing areas in the first row of a given sample group. When there is more than one value for a variable the values are averaged. Below are the collapsed rows for the above example.

sample	sample group	4	8	39	42
	BS-REF3:9/1/2005:SW:S:-:-	182	12.44	32	0.368
	EOF07:5/15/2005:SW:S:-:-	402	30	340	0.277
	EOF07:5/23/2005:SW:S:-:-			3000	1.397

REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

105. This Excel sheet is the summary spread sheet for the creation of the database. This spreadsheet (referred to as EDA_Sample in the CDM report) has 2,681 rows with 315 columns (each column is a variable).

106. This spreadsheet is further reduced by eliminating samples that have fewer than 20 observations on the 315 variables (i.e. of the 315 variables, only 19 or fewer have data). This results in an Excel sheet with 835 sample rows with at least 20 variable values in each row.

107. The database is further reduced by keeping only variables with < 24% of missing values. This produces an Excel sheet with 835 samples and 56 variables.

108. Finally, we retain only samples where all of the 56 variables have values (i.e. no missing values for the variables). This final Excel sheet has 419 sample rows with 56 variables and no missing values.

109.

<u>What I Did to Obtain the Maximum Available Data</u>	<u>Samples</u>	<u>Variables</u>
Full Data After Collapsing to Final Structure for Samples	2,681	315
<u>Samples with Less than 20 Variables with Data</u>	<u>1,846</u>	315
Samples with a Minimum of 20 Variables with Data	835	315
<u>Variables with data in less than 25% of samples</u>	835	<u>259</u>
Reduction to Variables with 25%+ Samples with Data	835	56
<u>Samples with Any Missing Data on 56 Variables</u>	<u>416</u>	56
Samples with No Missing Data on 56 Variables	419	56

<u>What Dr. Olsen Retained</u>		
Samples Dr. Olsen Retained	573	26
<u>Samples with Any Missing Data on 26 Variables</u>	<u>306</u>	26
Samples with No Missing Data on 26 Variables	267	26

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

110. In our reduction to the smallest dataset with no missing data, we obtained 56 variables. These 56 variables do NOT include the four bacteria variables. The Excel sheet created following Dr. Olsen's own methods is **missing four of the 26 variables** crucial to his PCA runs. The four missing variables are total coliforms, E. coli, enterococcus, and total coliforms.

111. The four bacteria variables were forced back into the analysis data set. This produced a large data set with 835 samples and 60 variables. However, when we then eliminate samples with missing data, we keep only **296 samples and 60 variables**. This is our final dataset for analysis, constructed considering only the use of all data and the elimination of samples with missing data.

112. There is no consistent explanation of the difference between Dr. Olsen's data and the data set we constructed. It is NOT accounted for with data rejected by Dr. Olsen because of his claims for samples in areas with cattle. There are, furthermore, severe differences between data on the Access database and data in Dr. Olsen's final Excel database, indicating that he: added data values in some cases with no documentation as to why, threw away data values in other cases, again with no documentation as to why, and on 3% of the records changed values with no explanation as to why.

113. Using the full set of data we derived with no missing values and including the bacterial data, we reanalyzed the data. Results are shown in Table 3 below.

REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Table 3. Analysis of the 296 Samples With 60 Variables.

Rotated	1	2	3	4	5	6	7	8
TOTAL_CADMIUM	0.973	0.065	-0.031	0.006	0.016	0.064	0.046	0.089
TOTAL_BERYLLIUM	0.962	0.131	-0.012	-0.005	0.023	0.044	0.091	0.032
TOTAL_SILVER	0.943	0.045	-0.014	0.04	0.038	0.06	-0.007	0.166
TOTAL_ANTIMONY	0.909	0.131	0.058	0.122	0.057	0.038	-0.067	-0.034
TOTAL_THALLIUM	0.894	0.095	0.059	0.099	0.073	0.032	-0.002	-0.001
DISSOLVED_CADMIUM	0.865	0.001	0.469	-0.077	-0.001	0.044	-0.043	-0.009
DISSOLVED_THALLIUM	0.858	0.007	0.49	-0.068	-0.008	0.042	-0.036	-0.012
DISSOLVED_BERYLLIUM	0.851	0.009	0.466	-0.061	0.012	-0.004	-0.032	-0.009
DISSOLVED_ALUMINUM	0.828	0.209	0.355	-0.208	0.003	0.093	0.07	0.042
TOTAL_SELENIUM	0.81	0.16	0.08	0.124	0.063	0.075	0.068	-0.22
TOTAL_KJELDAHL_NITROGEN	0.791	0.412	0.021	0.024	0.069	0.093	0.046	0.006
DISSOLVED_ANTIMONY	0.744	0.06	0.615	-0.062	0.034	0.066	-0.049	0.056
DISSOLVED_IRON	0.732	0.299	0.481	-0.196	0.067	0.1	0.038	0.005
DISSOLVED_LEAD	0.726	0.122	0.594	-0.069	0.004	0.041	0.001	-0.064
DISSOLVED_SILVER	0.725	0.033	0.666	-0.069	0.036	0.028	-0.047	0.033
DISSOLVED_VANADIUM	0.717	0.029	0.515	-0.014	-0.004	0.065	-0.027	0.244
DISSOLVED_COBALT	0.604	0.286	0.515	0.059	0.048	0.053	-0.01	0.076
TOTAL_P_4500PF_	0.129	0.883	0.171	0.076	-0.203	0.181	-0.029	0.042
TOTAL_COPPER	0.146	0.877	0.07	0.037	0.083	0.162	0.003	-0.063
TOC	0.138	0.851	0.149	0.046	0.143	0.242	0.002	-0.034
TOTAL DISSOLVED_P_4500PF_	0.067	0.815	0.183	0.142	-0.335	0.191	-0.139	0.033
TOTAL_POTASSIUM	0.121	0.814	0.207	0.367	-0.152	0.071	-0.059	0.022
TOTAL_NICKEL	0.268	0.783	0.11	0.176	0.082	0.045	0.235	0.266
SOLUBLE_REACTIVE_P_4500PF	0.034	0.783	0.166	0.129	-0.36	0.222	-0.093	0.059
TOTAL_ARSENIC	0.342	0.762	0.102	0.135	0.17	0.121	0.134	0.107
AMMONIA_NITROGEN	-0.257	0.759	0.12	0.061	0.227	0.118	0.094	-0.091
TOTAL_IRON	0.238	0.692	0.035	-0.338	0.255	0.146	0.398	0.02
TOTAL_ALUMINUM	0.234	0.665	0.026	-0.386	0.151	0.175	0.423	0.052
TOTAL_ZINC	0.513	0.622	0.052	-0.002	0.093	0.102	0.153	0.1
TOTAL_COBALT	0.572	0.611	0.015	-0.079	0.067	0.046	0.341	0.158
DISSOLVED_BARIUM	0.133	-0.164	0.913	0.091	-0.147	0.054	0.15	-0.005
DISSOLVED_MAGNESIUM	0.232	0.185	0.898	0.17	0.036	0.024	-0.046	-0.098
DISSOLVED_CALCIIUM	0.138	-0.292	0.861	0.287	-0.001	0.026	0.037	-0.038
DISSOLVED_CHROMIUM	-0.133	0.131	0.814	0.009	0.11	0	-0.055	-0.016
DISSOLVED_SODIUM	0.051	0.136	0.796	0.433	-0.165	-0.132	-0.166	0.118
DISSOLVED_SELENIUM	0.564	0.049	0.764	-0.079	0.03	0.039	-0.035	-0.155

REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

DISSOLVED_POTASSIUM	0.118	0.235	0.746	0.045	-0.221	-0.043	0.006	0.031
DISSOLVED_NICKEL	0.264	0.443	0.741	0.153	0.051	0.018	-0.065	0.193
DISSOLVED_ARSENIC	0.411	0.421	0.7	0.136	0.132	0.127	-0.059	0.045
DISSOLVED_COPPER	0.313	0.557	0.65	-0.084	0.025	0.138	-0.079	-0.091
DISSOLVED_ZINC	0.487	0.323	0.641	-0.017	-0.004	0.055	-0.097	0.02
DISSOLVED_MOLYBDENUM	0.6	0.178	0.606	0.036	0.029	0.015	-0.099	0.195
ALKALINITY_AS_CACO3	0.017	-0.208	-0.027	0.839	0.197	0.117	0.051	-0.144
TOTAL_CALCIIUM	-0.043	-0.396	-0.119	0.805	0.029	0.04	0.194	-0.046
TOTAL_SODIUM	-0.096	0.214	0.207	0.798	-0.215	-0.173	-0.179	0.19
CHLORIDE	-0.04	0.186	0.215	0.791	-0.235	-0.108	-0.141	0.188
TOTAL DISSOLVED SOLIDS	0.107	0.378	0.146	0.763	-0.071	0.006	0.117	0.056
TOTAL SULFATE_SO4	-0.106	0.343	0.215	0.676	-0.063	-0.162	-0.139	0.079
TOTAL_MANGANESE	0.104	0.581	0.118	-0.091	0.476	0.103	0.387	0.113
FECAL_COLIFORM	0.174	0.521	0.061	-0.058	0.039	0.772	0.053	0.042
E_COLI	0.102	0.533	0.045	-0.093	0.05	0.767	0.039	-0.022
TOTAL_COLIFORM	0.102	0.52	0.117	-0.023	0.016	0.713	0.032	-0.002
ENTEROCOCCUS_GROUP	0.192	0.502	-0.041	-0.094	0.057	0.688	0.109	0.122
TOTAL_BARIUM	-0.046	0.192	-0.19	0.068	-0.198	0.08	0.784	0.153
TOTAL_VANADIUM	0.112	0.103	0.038	0.121	0.143	0.065	0.134	0.822
TOTAL_LEAD	0.548	0.499	-0.009	-0.218	0.063	0.115	0.454	0.077
DISSOLVED_MANGANESE	0.183	0.305	0.599	0.024	0.41	0.051	0.157	0.056
NITRITE_NITRATE_AS_N	-0.185	-0.055	0.052	0.14	-0.774	-0.074	0.172	-0.136
TOTAL_MAGNESIUM	0.222	0.559	0.154	0.581	0.096	0.032	0.11	-0.142
TOTAL_CHROMIUM	-0.335	0.348	0.102	-0.012	0.141	-0.018	0.329	-0.144

"Variance" Explained by Rotated Components

1	2	3	4	5	6	7	8
14.67	11.987	10.655	5.131	1.92	2.72	2.025	1.339
5				7	3		

Percent of Total Variance Explained

1	2	3	4	5	6	7	8
24.45	19.978	17.758	8.552	3.21	4.53	3.376	2.231
8				2	8		

114. Many dissolved chemicals enter into the first and third principal components when they are included in the PCA runs. Phosphorus no longer loads onto the same component as fecal coliform, e-coli, total coliform and enterococcus. These latter do enter as a group to define a component, but not until the sixth principal component and not in the same component as the

REBUTTAL REPORT
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

phosphorus variables. These PCA results indicate that other variables are important in addition to the 26 variables discussed in the CDM report, and that the “signature” discovered by Dr. Olsen disappears when he brings in the full set of data available.